AD-A228 005

DTIC FILE COPY

(1)

Information and Computer Science

Working Notes of the

1990 Spring Symposium

on Automated Abduction

Paul O'Rorke

September 27, 1990

Technical Report 90-32

TECHNICAL REPORT

DTIC
ELECTE
OCT 18 1990
S D
Co D

UNIVERSITY OF CALIFORNIA
IRVINE

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

90 10 09 168

(1)

Working Notes of the 1990 Spring Symposium on Automated Abduction

Paul O'Rorke

September 27, 1990

Technical Report 90-32

DTIC
ELECTE
OCT 18 1990
S D
D

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

1990 Spring Symposium Series

Automated Abduction

Working Notes

DTIC COPY INSPECTED 6

Accession For
NTIS CRA&I ☑
DTIC TAB ☐
Unannounced ☐
Justification
By per call
Distribution /
Availability Codes
Dist | Avail and/or Special
A-1

March 27 – 29, 1990
Stanford University

STATEMENT "A" per Dr. Alan Meyrowitz
ONR/Code 1133IS
TELECON 10/16/90 VG

Sponsored by the
American Association for Artificial Intelligence
and
Office of Naval Research

List of Participants

Douglas E. Appelt
SRI International
333 Ravenswood Ave.
Menlo Park, CA 94025
APPELT@AI.SRI.COM

Lawrence Birnbaum
Inst. for the Learning Sciences
Dept. of Electrical Engineering
and Computer Science
Northwestern University
Evanston, IL 60208–9990
birnbaum@ils.nwu.edu
(708) 491-3500

Tom Bylander
Department of Computer
and Information Science
The Ohio State University
2036 Neil Avenue Mall
Columbus, OH 43210-1277
byland@cis.ohio-state.edu

Eugene Charniak
Department of Computer Science
Box 1910
Brown University
Providence, RI 02912

Steve A. Chien
Beckman Institute
405 North Mathews Avenue
University of Illinois
Urbana, IL 61801

William W. Cohen
AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974
wcohen@paul.rutgers.edu
wcohen@allegra.tempo.nj.att.com

Gregg Collins
The Institute for the
Learning Sciences
1890 Maple Ave
Evanston, IL 60201
(708) 491-3500
collins@ils.nwu.edu

Luca Console
Dipartimento di Informatica
Universita di Torino
Corso Svizzera 185
10149 Torino
ITALY
lconsole@ITOINFO.BITNET

M. J. Coombs
Computing Research Laboratory
New Mexico State University
Box 30001/3CRL
Las Cruces, NM 88003-0001

Andrea Danyluk
Department of Computer Science
Columbia University
New York, NY 10027
andrea@cs.columbia.edu
(212) 854-8121

Gerald DeJong
Beckman Institute
University of Illinois
405 North Mathews Street
Urbana, IL 61801
dejong@cs.uiuc.edu

Daniele Theseider Dupre'
Dipartimento di Informatica
Universita' di Torino
Corso Svizzera 185
10149 Torino
ITALY

Charles Elkan
Department of Computer Science
and Engineering C-014
University of California, San Diego
La Jolla, CA 92093
elkan@cs.ucsd.edu
(619) 534-1246

Brian Falkenhainer
Xerox Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto, CA 94304
falkenhainer.pa@xerox.com
(415) 494-4706

Tim Finin
Unisys Center for Advanced
Information Technology
P.O. Box 517
Paoli, PA 19301

Olivier Fischer
Laboratory for Artificial
Intelligence Research
The Ohio State University
2036 Neil Ave Mall
Columbus OH 43210.
fischer@cis.ohio-state.edu

Hector Geffner
IBM T.J.Watson Research Center
P. O. Box 704, Room H1-K10
Yorktown Heights, NY 10598
hector@ibm.com

Randy Goebel
Department of Computing Science
University of Alberta
Edmonton, Alberta
CANADA T6G 2H1
(403) 492-2683
goebel@cs.ualberta.ca

Robert P. Goldman
Computer Science Department
301 Stanley Thomas Hall
Tulane University
New Orleans, LA 70118-5698
rpg@rex.cs.tulane.edu
504-865-5840

Walter Hamscher
Price Waterhouse Tech. Centre
68 Willow Road
Menlo Park, CA 94025
Tel: (415) 688-6669
Fax: (415) 321-5543
E-mail: hamscher@tc.pw.com

Roger Hartley
Computing Research Laboratory
New Mexico State University
Box 30001/3CRL
Las Cruces, NM 88003-0001

Elizabeth Ann Hinkelman
Center for Information
and Language Studies
University of Chicago
1100 East 57th Street
Chicago, IL 60637
eliz@clove.uchicago.edu

Jerry Hobbs
Artificial Intelligence Center
SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025

Peter Jackson
McDonnell Douglas Research Lab
Dept. 225, Bldg. 105/2
Mailcode 1065165
P.O. Box 516
St. Louis, MO 63166
Tel: (314) 233 9271
Email: jackson@mdc.com

Dr. John R. Josephson
LAIR
Department of Computer
and Information Sciences
The Ohio State University
228 Bolz Hall
2036 Neil Avenue Mall
Columbus, OH 43210-1277
Netmail: jj@cis.ohio-state.edu

Kurt Konolige
SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025
konolige@warbucks.ai.sri.com

Bruce Krulwich
Institute for the Learning Sciences
Dept. of Electrical Engineering
and Computer Science
Northwestern University
Evanston, IL 60208–9990
krulwich@ils.nwu.edu
(708) 491-3500

Sridhar Mahadevan
IBM T.J.Watson Research Center
30 Saw Mill River Road
Hawthorne, NY 10532
sridhar@ibm.com

Raymond Mooney
Department of Computer Sciences
Taylor Hall 2.124
University of Texas at Austin
Austin, TX 78712

Steven Morris
Information and Computer Science
University of California, Irvine
Irvine, CA 92717
(714) 856-4840
morris@ics.uci.edu

Hwee Tou Ng
Department of Computer Sciences
Taylor Hall 2.124
University of Texas at Austin
Austin, TX 78712
htng@cs.utexas.edu

Peter Norvig
Computer Science Department
Evans Hall
University of California
Berkeley, CA 94720
Phone: (415) 642-9533
Fax: (415) 642-5775
Net: norvig@teak.berkeley.edu

Paul O'Rorke
Information and Computer Science
University of California, Irvine
Irvine, CA 92717
Phone: (714) 856-5563
Fax: (714) 856-4056
E-Mail: ororke@ics.uci.edu

Professor Judea Pearl
Engineering/ Computer Science
4731 Boelter Hall
UCLA
405 Hilgard Avenue
Los Angeles, CA 90024
pearl@cs.ucla.edu
(213) 825-3243

David Poole
Department of Computer Science
6356 Agricultural Road
University of British Columbia
Vancouver, B.C.
CANADA V6T1W5
poole@cs.ubc.ca
(604) 228-6254

Ashwin Ram
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332-0280
ashwin@cc.gatech.edu
(404) 853-9372

James A. Reggia
Department of Computer Science
Institute for Advanced
Computer Studies
University of Maryland
College Park, MD 20742
reggia@cs.UMD.EDU

Raymond Reiter
Department of Computer Science
University of Toronto
Toronto, Ontario
CANADA M5S 1A4

Roger Schank
Institute for the Learning Sciences
Northwestern University
1890 Maple Avenue
Evanston, IL 60201

Bart Selman
Department of Computer Science
University of Toronto
Toronto, Ontario
CANADA M5S 1A4
bart@ai.toronto.edu
(416) 978-5182

Murray Shanahan
Department of Computing
Imperial College
Science, Technology and Medicine
180 Queen's Gate
London SW7 2BZ
ENGLAND

Jack W. Smith
The Ohio State University
Laboratory for Knowledge-Based
Medical Systems
571 Health Science Library
376 W. 10th Ave.
Columbus, OH 43210
jsmith@osu-20.ircc.ohio-state.edu
jsmith@cis.ohio-state.edu

Mark E. Stickel
Artificial Intelligence Center
SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025
stickel@ai.sri.com
(415) 859-6151

Paul Thagard
Princeton University
Cognitive Science Laboratory
221 Nassau Street
Princeton, NJ 08542
pault@clarity.princeton.edu

Pietro Torasso
Dipartimento di Informatica
Universita di Torino
Corso Svizzera 185
10149 Torino
ITALY

Robert Wilensky
Computer Science Department
Evans Hall
University of California
Berkeley, CA 94720
wilensky@larch.berkeley.edu

Table of Contents

Theory

Tasks

FOREWORD

The philosopher Charles Sanders Peirce used the term abduction for a form of inference considered to be as important as deduction and induction (Peirce, 1931-1958). His description of abduction was basically: "The surprising fact C is observed; But if A were true, C would be a matter of course, hence there is reason to suspect that A is true." Peirce's abduction replaced his earlier theory of the "method of hypothesis" (Thagard, 1981). Abduction is concerned with explanatory reasoning and is closely related to the relatively modern notions of "backward chaining" and "inference to the best explanation" (Harman, 1965; Josephson, 1990).

Since explanations are important in many different aspects of intelligence, cognitive scientists have become interested in computer programs that construct and evaluate explanations. In artificial intelligence, a number of key tasks have come to be viewed in terms of abduction. In expert systems, the best known abduction problem is diagnosis. In natural language comprehension, plan recognition is viewed as an abduction problem involving the inference of goals from observed behavior. In qualitative physics, postdiction is an abduction problem involving explaining states of the physical world in terms of processes and causal laws. In machine learning, explanation-based learning (EBL) strategies improve performance using processes that construct explanations.

Abduction-related work has been done in different areas of AI for nearly twenty years (Pople, 1973), but until recently researchers working in different subfields often failed to recognize that they might benefit from work on abduction by people in other areas. The spring symposium on Automated Abduction, sponsored by AAAI and ONR and held at Stanford in March of 1990, aimed to facilitate cross-fertilization in the hope of accelerating research advances in all subfields of AI concerned with explanations.

Researchers with interests in business, planning, diagnosis, qualitative physics, machine learning and discovery, and natural language processing gathered to discuss the role of abduction in their disciplines. Walter Hamscher pointed out the potential for applications in business and introduced a system named after Sherlock Holmes's banker. Hamscher's CROSBY, based upon de Kleer and Williams's model-based diagnosis program SHERLOCK, automatically constructs plausible explanations for unexpected financial results. Charles Elkan's contribution describe an approach to planning using abductive assumptions to generate approximate, incremental plans (see also Elkan, 1990). Bruce Krulwich, Lawrence Birnbaum, and Gregg Collins described a goal directed approach to learning strategic concepts from expectation failures during plan execution (for related work, see Birnbaum, Collins, Freed, & Krulwich, 1990). Murray Shanahan presented abductive solutions to temporal projection problems such as Henry Kautz's stolen car problem and the bloodless version of the Yale shooting problem due to Hanks and McDermott.

Robert Goldman and Eugene Charniak began a session on abduction and natural language understanding by presenting their work as a special case of a general probabilistic approach to abduction (Charniak & Shimony, 1990). Mark Stickel described a general logic and cost-based approach to abduction, and Jerry Hobbs provided an integrated approach to

natural language processing and discourse interpretation based upon this abduction method. Elizabeth Hinkelman described her recent thesis work on abductive speech act recognition. Ashwin Ram sketched his recent thesis work on a program called AQUA, which builds explanations in order to find answers to questions that arise in the process of text comprehension (see Ram, 1989). Hwee Tou Ng and Raymond Mooney discussed the role of explanatory coherence in natural language interpretation and observed that "Occam's Razor isn't sharp enough" (Ng & Mooney, 1989, 1990). Preferring maximally general explanations doesn't always work well. Coherence seems to be more important than generality. Peter Norvig and Robert Wilensky pointed out some weaknesses of the current abductive approaches to NLP based upon coherence, cost, and probability, listing a number of problems that still need to be addressed in constructing general abductive models of comprehension.

In a session on abduction and learning, I argued that progress in research on abduction can be used to improve our ideas about explanation-based learning (EBL). In particular, I argued that replacing the theorem provers traditionally used to construct explanations in EBL with abduction engines enables EBL systems to deal with the conflicting plausible explanations that arise when theories are incomplete or incorrect. Furthermore, abductive inference makes it possible for EBL systems to learn at the knowledge level (O'Rorke, 1988, 1990). Steven Morris presented an approach to theory revision using abduction for hypothesis formation and illustrated the potential for learning at the knowledge level using an example based on the chemical revolution (O'Rorke, Morris, & Schulenburg, 1990). Bill Cohen presented another approach to revising imperfect theories using abductive EBL. He illustrated the performance of his method on the problem of learning the concept "good opening bid" in the card game bridge (see also Cohen, 1989, 1990). Sridhar Mahadevan presented a technique for acquiring rules that extend incomplete theories containing "determinations." Andrea Danyluk discussed the importance of contextual knowledge in constructing explanations for EBL. She described experiments from her thesis work testing her methods in network fault diagnosis domains. Steve Chien presented results from his thesis work on EBL for incremental, approximate planning. Gerald DeJong, the originator of EBL (DeJong, 1988; DeJong & Mooney, 1986) argued that narrow conceptions of abduction do not provide the kind of plausible inference necessary for explanation-based learning based upon imperfect knowledge.

The workshop included sessions on task independent methods and general theories of abduction. A number of people working on tasks such as diagnosis and natural language comprehension have devoted considerable time to the development of domain-independent methods for constructing and evaluating explanations. I have already mentioned general approaches to abduction arising out of work on NLP based upon logic, cost minimization, coherence, "explanation patterns," and probability. Collaborative work such as the work of Harry Pople and Jack Myers has led to insights into medical diagnosis as an abduction problem. Olivier Fischer and Jack Smith contributed an analysis and comparison of INTERNIST and the RED-2 system developed by computer scientists and M.D.s at Ohio State. The generalized set covering and parsimonious covering theories (PCT) of James Reggia and Yun Peng were also inspired by work on diagnostic problem solving. Reggia sketched PCT (elaborated fully in Peng & Reggia, 1990), described applications of PCT to non-diagnostic tasks such as natural language processing, and compared PCT to a general theory of explanatory coherence

(TEC) presented by Paul Thagard. Thagard showed how TEC could be used to provide a model of decision making that can be implemented in connectionist networks. Roger Hartley and Michael Coombs's contribution described an architecture called MGR extending generalized set covering methods for abduction to model generative reasoning. Raymond Reiter sketched abduction related work arising out of research on model-based diagnosis and the foundations of assumption based truth maintenance systems and emphasized the importance of computing prime implicants (see also Reiter, 1987; Reiter & de Kleer, 1987; de Kleer, Mackworth, & Reiter 1990).

Several general, relatively formal logical approaches to abduction were discussed. David Poole sketched his THEORIST system for abduction, relating it to default and hypothetico-deductive reasoning and discussing its application to design. Douglas Appelt presented initial work on using model preferences to generalize existing approaches to abduction based upon Bayesian probability, minimizing abnormality, and maximizing defaults. Kurt Konolige presented a theory which included a general framework for abduction and which clarified the relationship between abduction and diagnostic reasoning methods using closure, minimization, and consistency (see also Junker & Konolige, 1990). Luca Console, Daniele Dupre, and Pietro Torasso described related work providing a semantics for abduction and clarifying the relationship between abduction and deductive reasoning. Peter Jackson also provided a semantic account of abductive inference and showed how it can be done in terms of counterfactual reasoning if completeness assumptions are introduced.

In addition to axiomatic characterizations of abduction and semantic theories of abduction, several analyses of the complexity of abductive computations were presented. It is probably not surprising that abduction, like many other AI problems, is intractable in general, but interesting results were presented by Tom Bylander and Bart Selman which more exactly characterize when and why abduction is hard (see also Selman & Levesque, 1990). It was encouraging to see that several general formal theories of abduction have begun to develop and more encouraging to see these theories tied closely to each other and to the algorithms being used in applications.

Lively discussions comparing different approaches, methods and implementations (e.g., Bayesian probabilistic reasoning vs. connectionist networks) took place both on and off-line. These discussions were sometimes quite heated. At one point I was asked why I had invited a certain speaker since it was "like inviting a creationist to a scientific meeting." Another participant wanted to know why Judea Pearl and Roger Schank were invited to give "back-to-back" talks presenting their views of abduction.

Roger Schank gave an invited talk encouraging workshop participants to spend more time on memory-based approaches to explanation and less time on approaches based upon problem solving, theorem proving and probability theory. Schank's views are described more fully in his book "Explanation Patterns" (Schank, 1986). In his invited talk, Judea Pearl made a strong case for a general probabilistic approach to abduction. He described probabilistic methods for defining the primitive causal relationships underlying theories of causal explanation. In addition, his submission discussed the relationship between probabilistic and qualitative approaches to abduction. Pearl's views are stated more fully in Pearl (1988).

The sharply contrasting invited talks were a result of scheduling constraints, but they

highlighted the differences in the points of view of the participants. One participant stalked out at one point, informing me that he had "had it up to here" with the logicists' view of abduction. A surprisingly heated exchange occurred between a logicist and a cognitive scientist. On the whole, however, the differences between participants were expressed in friendly and valuable constructive criticism.

Relationships between abduction and various forms of inference were explored by a number of participants. Brian Falkenhainer discussed analogical reasoning and argued in his contribution that deduction, abduction, and analogy are closely related. John Josephson characterized abduction in terms of an inference schema related to Harman's notion of inference to the best explanation. Josephson also discussed the logical form of abduction and its relationship to deduction and induction. Randy Goebel described abduction as a "logical method of isolating interesting hypotheses" and discussed its relationships to hypothetico-deductive reasoning, deduction, induction, analogy, probability, and non-monotonic reasoning. Hector Geffner focused on the relationship to default reasoning and described a special class of default theories using modal causal operators (Geffner, 1990).

The workshop provided a broad overview of the rapidly accumulating work on abduction and brought together a number of researchers who ordinarily operate in disjoint subfields of AI. Many participants found the technical exchanges and the discussions of relationships very valuable. If I were forced to identify weaknesses of the workshop, I would admit to the fact that little or no psychological data was presented about how people construct and evaluate explanations and few formal evaluations or comparisons of alternative approaches or systems were given. While there may be conferences on abduction in the future which will put a stronger emphasis on evaluation, the quality of the work was quite high for a workshop. The symposium provided a useful snapshot of an important, fundamental research area emerging at the intersection of several subfields of AI.

My thanks to Hector Levesque, Carol Hamilton, and AAAI for making the symposium possible. Thanks also to the other organizers and members of the program committee: Eugene Charniak, Gerald DeJong, Jerry Hobbs, Jim Reggia, Roger Schank, and Paul Thagard. Caroline Ehrlich and Steven Morris helped with preparations at UCI. AAAI and Alan Meyrowitz of the AI Program in the Office of Naval Research generously provided travel support enabling graduate students to participate. My apologies if I have tread on anyone's toes. I am happy to accept corrections of any errors I may have made in giving my impressions of the workshop.

AAAI policy limits distribution of symposium working notes to attendees. However, with encouragement from AAAI and SIGART and with permission of the authors (who retain copyrights) the working notes of the abduction symposium are now available as a UCI technical report. Enjoy!

References

Birnbaum, L., Collins, G., Freed, M., & Krulwich, B. (1990). Model-based diagnosis of planning failures. *The Eighth National Conference on Artificial Intelligence* (pp. 318–323). Boston, MA: AAAI Press/ The MIT Press.

Charniak, E., & Shimony, S. E. (1990). Probabilistic semantics for cost based abduction. *The Eighth National Conference on Artificial Intelligence* (pp. 106–111). Boston, MA: AAAI Press/ The MIT Press.

Cohen, W. W. (1989). *Abductive explanation-based learning: A solution to the multiple explanation problem* (Technical Report). New Brunswick, NJ: Rutgers University, Department of Computer Science.

Cohen, W. W. (1990). Learning from textbook knowledge: A case study. *The Eighth National Conference on Artificial Intelligence* (pp. 743-748). Boston, MA: AAAI Press/ The MIT Press.

DeJong, G. F. (1988). An introduction to explanation-based learning. In H. Shrobe (Ed.), *Exploring artificial intelligence.* San Mateo, CA: Morgan Kaufmann.

DeJong, G. F., & Mooney, R. (1986). Explanation-based learning: An alternative view. *Machine Learning, 1*, 145–176.

de Kleer, J., Mackworth, A. K., & Reiter, R. (1990). Characterizing diagnoses. *The Eighth National Conference on Artificial Intelligence* (pp. 324–331). Boston, MA: AAAI Press/ The MIT Press.

Elkan, C. (1990). Incremental, approximate planning. *The Eighth National Conference on Artificial Intelligence* (pp. 145–150). Boston, MA: AAAI Press/ The MIT Press.

Geffner, H. (1990). Causal theories for nonmonotonic reasoning. *The Eighth National Conference on Artificial Intelligence* (pp. 524–530). Boston, MA: AAAI Press/ The MIT Press.

Harman, G. (1965). The inference to the best explanation. *Philosophical Review, 74*, 88–95.

Josephson, J. R. (1990). On the "logical form" of abduction. *Working Notes of the AAAI Spring Symposium on Automated Abduction* (Technical Report 90-32). Irvine: University of California, Department of Information and Computer Science.

Junker, U., & Konolige, K. (1990). Computing the extensions of autoepistemic and default logics with a truth maintenance system. *The Eighth National Conference on Artificial Intelligence* (pp. 278–283). Boston, MA: AAAI Press/ The MIT Press.

Ng, H. T., & Mooney, R. (1990). On the role of coherence in abductive explanation. *The Eighth National Conference on Artificial Intelligence* (pp. 337–342). Boston, MA: AAAI Press/ The MIT Press.

Ng, H. T., & Mooney, R. J. (1989). Occam's Razor isn't sharp enough: The importance of coherence in abductive explanation. *Proceedings of the Second AAAI Workshop on Plan Recognition.* Detroit, MI.

O'Rorke, P. (1988). Automated abduction and machine learning. In G. DeJong (Ed.), *Working Notes of the AAAI Spring Symposium on Explanation-Based Learning* (pp. 170–174). Palo Alto, CA.

O'Rorke, P., Morris, S., & Schulenburg, D. (1989). Abduction and world model revision. *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society.* Ann Arbor, MI: Lawrence Erlbaum. An expanded version appears as "Theory formation by abduction: A case study based on the chemical revolution." In J. Shrager & P. Langley (Eds.), *Computational models of scientific discovery and theory formation* (pp. 197–224). San Mateo, CA: Morgan Kaufmann.

O'Rorke, P. (1990). Integrating abduction and learning. *Working Notes of the AAAI Spring Symposium on Automated Abduction* (Technical Report 90-32). Irvine: University of California, Department of Information and Computer Science.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* San Mateo, CA: Morgan Kaufmann.

Peirce, C. S. S. (1931-1958). *Collected papers of Charles Sanders Peirce (1839-1914).* Cambridge, MA: Harvard University Press.

Peng, Y., & Reggia, J. A. (1990). *Abductive inference models for diagnostic problem solving.* New York: Springer-Verlag.

Pople, H. E. (1973). On the mechanization of abductive logic. *Proceedings of the Third International Joint Conference on Artificial Intelligence* (pp. 147–152). Stanford, CA: Morgan Kaufmann.

Ram, A. (1989). *Question-driven understanding: An integrated theory of story understanding, memory, and learning.* Doctoral dissertation, Yale University, New Haven, CT.

Reiter, R. (1987). A theory of diagnosis from first principles. *Artificial Intelligence, 32,* 57–95.

Reiter, R., & de Kleer, J. (1987). Foundations of assumption-based truth maintenance systems: Preliminary report. *Proceedings of the Sixth National Conference on Artificial Intelligence* (pp. 183–188). Seattle, WA: Morgan Kaufmann.

Schank, R. C. (1986). *Explanation patterns: Understanding mechanically and creatively.* Hillsdale, NJ: Lawrence Erlbaum.

Selman, B., & Levesque, H. J. (1990). Abductive and default reasoning: A computational core. *Proceedings of the Eighth National Conference on Artificial Intelligence* (pp. 343–348). Boston, MA: AAAI Press/ The MIT Press.

Thagard, P. R. (1981). *Peirce on hypothesis and abduction.* Lubbock, TX: Texas Tech University Press.

Incremental Construction of Probabilistic Models for Language Abduction Work in Progress

Robert P. Goldman and Eugene Charniak*

Dept. of Computer Science, Brown University
Box 1910,
Providence, RI 02912

For some time we have been interested in the problems posed by uncertainty in story understanding. Our particular interest is in plan-recognition as it is needed in text understanding: understanding the meanings of stories by understanding the way the actions of characters in the story serve purposes in their plans. Our work builds upon earlier work in script- and plan-based understanding of stories like that of Wilensky [1983], Charniak [1986] and Norvig [1987].

We see plan-recognition and text understanding as a particular case of the problem of abduction.[1] In particular, for the case of simple, declarative text, we view the language user as a transducer. The language user observes some thing (event or object) in the 'real world', and translates this thing into language. Our task is to reason from the text to the intentions of the language user and thence to the thing described.

Because abduction problems involve uncertainty, we have adopted a probabilistic approach to the problem of story comprehension. In order to make such an approach feasible, a number of techniques have been used:

1. Simplifying assumptions
2. A graphical representation of the probabilistic model
3. Incremental construction and evaluation of the representation of this probabilistic model.

We represent the plans in an isa-hierarchy of frames. We assume that this set of plans is exhaustive. We also assume that actions are related to each other only through the mediation of these plans. So, for example, in reading stories about various people's day-to-day activities, we will make inferences about an agent's travel based on plans to achieve everyday tasks. We would *not* take into account a particular person's systematic preference for walking, rather than driving. For simple stories this does not seem to be a problem.

We have chosen to represent the resulting probabilistic inference problem using belief networks.[2] Belief networks are directed acyclic graphs that can be used to represent probability problems. In a belief network, nodes represent random variables, and arcs represent direct influences between random variables. There are three advantages to belief networks as representations for probability distributions. First of all, properties of conditional independence can be read off a belief network. Second, the probability distribution corresponding to a belief network may be represented locally. For each node, it suffices to provide a conditional probability distribution for each combination of values of its parent nodes. Finally, while in general the problem of determining the posterior distribution of a partially-instantiated belief network is NP-hard [Cooper, 1987], considerable attention has been devoted to finding efficient approaches to evaluating such networks.

A sample belief network for the story "Jack got a rope. He killed himself." is given as Figure 1. The nodes at the bottom represent the evidence, we have observed: three words, "kill", "get" and "rope" and

*This work has been supported in part by the National Science Foundation under grants IST 8416034 and IST 8515005 and Office of Naval Research under grant N00014-79-C-0529.

[1] See [Hobbs *et al.*, 1988] for a statement of this position.

[2] Judea Pearl's book [Pearl, 1988] gives a thorough account of the properties of such networks.

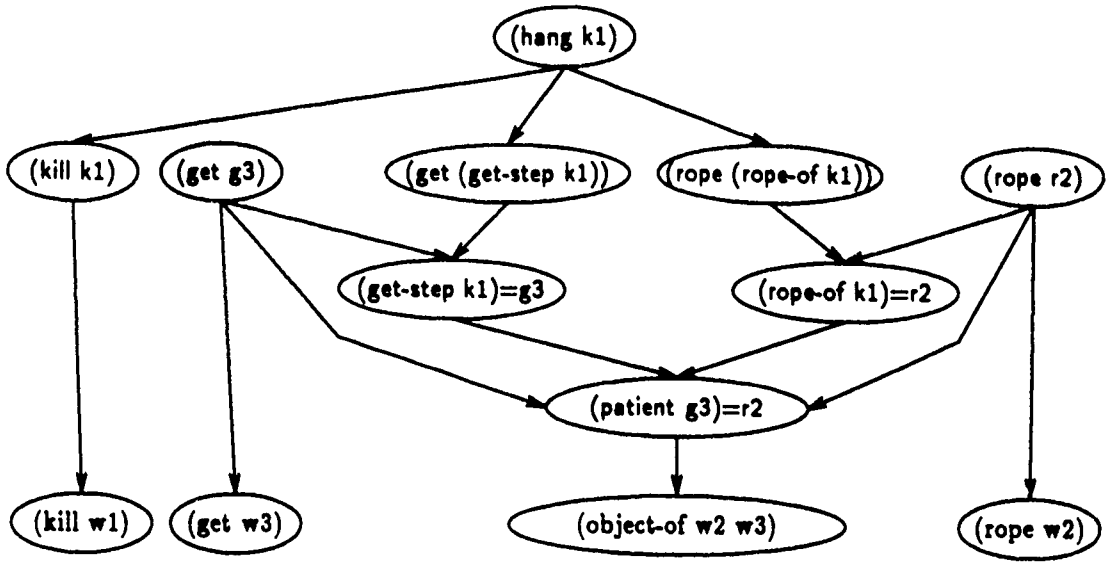


Figure 1: The Bayesian network for "Jack got a rope. He killed himself."

the fact that the rope is the object of the get. Nodes with arcs into the words represent possible causes for the use of these words. E.g., one possible cause for using the word "rope" is that the author wishes to talk about a rope: (rope r2). One reason for the rope being the object of the verb "get" is that it is the patient of the getting action the word refers to. If the kill referred to were a hanging, that would dictate the presence of a getting action whose patient is a rope. The getting action and the rope we've postulated might fill those roles (the equality statements).

This example is simplified for the purpose of clarity, showing only one possible interpretation for the input. The actual diagrams used in our program are more complicated. This figure also illustrates how the posterior distribution over a belief net can determine the interpretation of a text. We are concerned with the probability that Jack has a plan to hang himself, given the input we have observed. I.e., we are interested in

$$P((\texttt{hang k1}) \mid (\texttt{kill w1}), (\texttt{get w3}), (\texttt{object w2 w3}), (\texttt{rope w2}))$$

Because a full probabilistic model for any utterance might be infinite, we construct and evaluate only small pieces of this model at any given time. We have developed a language for writing network-building rules, and a set of such rules for our domain. These rules are similar to the forward-chaining rules in a conventional TMS. They differ in that rules are not restricted to adding justifications to some derived statement. In fact, since we are trying to build networks for a diagnostic problem, our rules will typically be triggered by the heads of arcs they add, rather than by the tails. Our network-building rules also provide more information than simple connectivity: they contain information used to compute the conditional probability matrices of the nodes in the network.

The technique of incrementally building belief networks with production rules may be of more general applicability than language abduction. However, several features of this domain make this technique particularly appropriate. Because our program is passive, it cannot seek out new evidence. This helps us focus our search: we are always going to be searching from evidence to explanations. We do not need to look for new observations. Our domain also makes it possible to build up a model out of nodes of stylized types which can be parameterized. This approach was suggested by Pearl [1988], who suggested "noisy-or gates". We use these as well as noisy XORs, ORs and ANDs. For example, our equality statements are parameterized noisy-ANDs: in order for the equality to be possible, both terms must be of the same type (e.g., in Figure 1, both g3 and (get-step k1) must be gets in order for (get-step k1) = g3 to be possible.

There has been similar work contemporary with ours. [Breese, 1989], and [D'Ambrosio, 1988] de-

scribe techniques for constructing models on an as-needed basis. [Levitt *et al.*, 1989] discusses incremental model evaluation and extension.

This approach is being tested in a program called Wimp3, which works as follows:

1. A parser reads one word of the English text. It produces statements which describe the words in the story and the syntactic relations between them.

2. The output of the parser is taken by the network construction component. This component contains rules for language abduction. It builds a net, or extends the current net if some input has already been received.

3. The resulting belief network is evaluated by a network-evaluation component. If certain conclusions are overwhelmingly favored, they may be accepted as true to simplify further computation.

4. Return to step 1.

This work is more fully described in three other papers of the authors: [Charniak and Goldman, 1989b] and [Charniak and Goldman, 1989a] give a probabilistic account of the problem of story understanding. This provides the mathematical foundation of this work. [Goldman and Charniak, 1989] gives a more detailed account of the program. Finally, [Carroll and Charniak, 1989] discusses a marker-passer which is used to control the search of the belief network construction rules. Two other papers, [Goldman and Charniak, 1988] and [Charniak and Goldman, 1988] describe the authors' earlier work in this area, an attempt to build a logical-probabilistic hybrid program for story understanding.

References

[Breese, 1989] John S. Breese. Construction of belief and decision networks. forthcoming, 1989.

[Carroll and Charniak, 1989] Glenn Carroll and Eugene Charniak. Finding plans with a marker-passer. In *Proceedings of the Plan-recognition workshop*, 1989.

[Charniak and Goldman, 1988]
Eugene Charniak and Robert P. Goldman. A logic for semantic interpretation. In *Proceedings of the Annual Meeting of the ACL*, 1988.

[Charniak and Goldman, 1989a]
Eugene Charniak and Robert P. Goldman. Plan recognition in stories and in life. In *Proceedings Workshop on Uncertainty and Probability in AI.* Morgan Kaufmann Publishers, Inc., 1989.

[Charniak and Goldman, 1989b] Eugene Charniak and Robert P. Goldman. A semantics for probabilistic quantifier-free first-order languages, with particular application to story understanding. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence*, 1989.

[Charniak, 1986] Eugene Charniak. A neat theory of marker passing. In *Proceedings of the Sixth National Conference on Artificial Intelligence*, pages 584–589, 1986.

[Cooper, 1987] Gregory F. Cooper. Probabilistic inference using belief networks is np-hard. Technical Report KSL-87-27, Medical Computer Science Group, Stanford University, 1987.

[D'Ambrosio, 1988] Bruce D'Ambrosio. Process, structure and modularity in reasoning with uncertainty. In *The Fourth Workshop on Uncertainty in Artificial Intelligence*, pages 64–72, 1988.

[Goldman and Charniak, 1988] Robert P. Goldman and Eugene Charniak. A probabilistic ATMS for plan recognition. In *Proceedings of the Plan-recognition workshop*, 1988.

[Goldman and Charniak, 1989] Robert P. Goldman and Eugene Charniak. A probabilistic approach to plan recognition and text understanding: Work in progress. In *Proceedings of the Plan-recognition workshop*, 1989.

[Hobbs *et al.*, 1988] Jerry R. Hobbs, Mark Stickel, Paul Martin, and Douglas Edwards. Interpretation as abduction. In *Proceedings of the 26th Annual Meeting of the ACL*, pages 95–103, 1988.

[Levitt *et al.*, 1989] T.S. Levitt, J.M. Agosta, and T.O. Binford. Model-based influence diagrams for machine vision. In *Proceedings Workshop on Uncertainty and Probability in AI*, pages 233–244. Morgan Kaufmann Publishers, Inc., 1989.

[Norvig, 1987] Peter Norvig. Inference in text understanding. In *Proceedings of the Seventh National Conference on Artificial Intelligence*, pages 561–565, 1987.

[Pearl, 1988] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, Inc., 95 First Street, Los Altos, CA 94022, 1988.

[Wilensky, 1983] Robert Wilensky. *Planning and Understanding*. Addison-Wesley, Reading, Mass., 1983.

A Method for Abductive Reasoning in Natural-Language Interpretation

Mark E. Stickel

Artificial Intelligence Center
SRI International
Menlo Park, California 94025

Introduction

Abductive inference is inference to the best explanation. The process of interpreting sentences in discourse can be viewed as the process of generating the best explanation as to why a sentence is true, given what is already known [3]; this includes determining what information must be added to the listener's knowledge (what assumptions must be made) for the listener to know the sentence to be true. Some new forms of abduction are more appropriate to the task of interpreting natural language than those used in the traditional diagnostic and design synthesis applications of abduction. In one new form, least specific abduction, only literals in the logical form of the sentence can be assumed. The assignment of numeric costs to axioms and assumable literals permits specification of preferences on different abductive explanations. Least specific abduction is sometimes too restrictive. Better explanations can sometimes be found if literals obtained by backward chaining can also be assumed. Assumption costs for such literals are determined by the assumption costs of literals in the logical form and functions attached to the antecedents of the implications. There is a new Prolog-like inference system that computes minimum-cost explanations for these abductive reasoning methods.

We consider here the abductive explanation of conjunctions of positive literals from Horn clause knowledge bases. An explanation will consist of a substitution for variables in the conjunction and a set of literals to be assumed. In short, we are developing an abductive extension of pure Prolog.

*This abstract is condensed from Stickel [7]. The research was supported by the Defense Advanced Research Projects Agency, under Contract N00014-85-C-0013 with the Office of Naval Research, and by the National Science Foundation, under Grant CCR-8611116. The views and conclusions contained herein are those of the author and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency, the National Science Foundation, or the United States government. Approved for public release. Distribution unlimited.

Four Abduction Schemes

In general, if the formula $Q_1 \wedge \cdots \wedge Q_n$ is to be explained or abductively proved, the substitution θ and the assumptions $P_1, \ldots, P_m$ would constitute one possible explanation if $(P_1 \wedge \cdots \wedge P_m) \supset (Q_1 \wedge \cdots \wedge Q_n)\theta$ is a consequence of the knowledge base.

It is a general requirement that the conjunction of all assumptions made be consistent with the knowledge base. With an added factoring operation and without the literal ordering restriction, so that any, not just the leftmost, literal of a clause can be resolved on, Prolog-style backward chaining is capable of generating all possible explanations that are consistent with the knowledge base. That is, every possible explanation consistent with the knowledge base is subsumed by an explanation that is generable by backward chaining and factoring. It would be desirable if the procedure were guaranteed to generate no explanations that are inconsistent with the knowledge base, but this is impossible.

Obviously, any clause derived by backward chaining and factoring can be used as a list of assumptions to prove the correspondingly instantiated initial formula abductively. This can result in an overwhelming number of possible explanations. Various abductive schemes have been developed to limit the number of acceptable explanations. These schemes differ in their specification of which literals are assumable.

What we shall call *most specific abduction* has been used particularly in diagnostic tasks [4,1]. In explaining symptoms in a diagnostic task, the objective is to identify causes that, if assumed to exist, would result in the symptoms. The most specific causes are usually sought, since identifying less specific causes may not be as useful. In most specific abduction, the only literals that can be assumed are those to which backward chaining can no longer be applied.

What we shall call *predicate specific abduction* has been used particularly in planning and design synthesis tasks [2]. In generating a plan or design by specifying its objectives and ascertaining what assumptions must be made to make the objectives provable, acceptable assumptions are often expressed in terms of a prespecified set of predicates. In planning, for example, these might represent the set of executable actions.

The criterion for "best explanation" used in natural-language interpretation differs greatly from that used in most specific abduction for diagnostic tasks. To interpret the sentence "the watch is broken," the conclusion will likely be that we should add to our knowledge the information that the watch currently discussed is broken. The explanation that would be frivolous and unhelpful in a diagnostic task is just right for sentence interpretation. A more specific causal explanation, such as a broken mainspring, would be gratuitous.

Predicate specific abduction is not ideal for natural-language interpretation either, since there is no easy division of predicates into assumable and nonassumable, so that those assumptions that can be made will be reasonably restricted. Most predicates must be assumable in some circumstances such as when certain sentences are being interpreted, but in many other cases should not be assumed.

As an alternative, we consider what we will call *least specific abduction* to be well suited to natural-language-interpretation tasks. It allows only literals in the initial formula to be assumed and thereby seeks to discover the least specific assumptions that explain a sentence. More specific explanations would unnecessarily and often incorrectly require excessively detailed assumptions.

We note that assuming any literals other than those in the initial formula generally results in more specific and thus more risky assumptions. When explaining R with $P \supset R$ (or $P \wedge Q \supset R$) in the knowledge base, either R or P (or P and Q) can be assumed to explain R. Assumption of R, the consequent of an implication, in preference to the antecedent P (or P and Q), results in the fewest consequences.

Although least specific abduction is often sufficient for natural-language interpretation, it is clearly sometimes necessary to assume literals that are not in the initial formula. We propose *chained specific abduction* for these situations. Assumability is inherited—a literal can be assumed if it is an assumable literal in the initial formula or if it can be obtained by backward chaining from an assumable literal.

Factoring some literals obtained by backward chaining and assuming the remaining antecedent literals can also sometimes yield better explanations. When $Q \wedge R$ is explained from

$$P_1 \wedge P_2 \supset Q$$
$$P_2 \wedge P_3 \supset R$$

the explanation that assumes P_1, P_2, and P_3 may be preferable to the one that assumes Q and R. Even if Q and R are not provable, it might not be necessary to assume all of P_1, P_2, and P_3, since some may be provable.

Assumption Costs

A key issue in abductive reasoning is picking the best explanation. Defining this is so subjective and task dependent that there is no hope of devising an algorithm that will always compute only the best explanation. Nevertheless, there are often so many abductive explanations that it is necessary to have some means of eliminating most of them. We attach numeric assumption costs to assumable literals, and compute minimum-cost abductive explanations in an effort to influence the abductive reasoning system toward favoring the intended explanations.

We regard the assignment of numeric costs as a part of programming the explanation task. The values used may be determined by subjective estimates of the likelihood of various interpretations, or perhaps they may be learned through exposure to a large set of examples.

If only the cost of assuming literals is counted in the cost of an explanation, there is in general no effective procedure for computing a minimum-cost explanation. For example, if we are to explain P, where P is assumable with cost 10, then assuming P produces an explanation with cost 10, but proving P would result in a better explanation with cost 0. Since provability is undecidable in general, it may be impossible to determine whether the cost 10 explanation is best.

The solution is that the cost of proving literals must also be included in the cost of an explanation. An explanation that assumes P with cost 10 would be preferred to an explanation that proves P with cost 50 (e.g., in a proof of 50 steps) but would be rejected in favor of an explanation that proves P with cost less than 10.

There are substantial advantages gained by taking into account proof costs as well as assumption costs, in addition to the crucial benefit of making theoretically possible the search for a minimum-cost explanation.

If costs are associated with the axioms in the knowledge base as well as with assumable literals, these costs can be used to encode information on the likely relevance of the fact or rule to the situation in which the sentence is being interpreted.

We have some reservations about choosing explanations on the basis of numeric costs. Nonnumeric specification of preferences is an important research topic. Nevertheless, we have found these numeric costs to be quite practical; they offer an easy way of specifying that one literal is to be assumed rather than another. When many alternative explanations are possible, summing numeric costs in each explanation, and adopting an explanation with minimum total cost, provides a mechanism for comparing the costs of one proof and set of assumptions against the costs of another. If this method of choosing explanations is too simple, other means may be too complex to be realizable. We provide a procedure for computing a minimum-cost explanation by enumerating possible partial explanations in order of increasing cost. Even a perfect scheme for specifying preferences among alternative explanations may not lead to an effective procedure for generating a most preferred one. Finally, any scheme will be imperfect: people may disagree as to the best explanation of some data and, moreover, sometimes do misinterpret sentences.

Minimum-Cost Proofs

We now present the inference system for computing abductive explanations. This method applies to predicate specific, least specific, and chained specific abduction.

Every literal Q_i in the initial formula is annotated with its assumption cost c_i:

$$Q_1^{c_1}, \ldots, Q_n^{c_n}$$

The cost c_i must be nonnegative; it can be infinite, if Q_i is not to be assumed.

Every literal P_j in the antecedent of an implication in the knowledge base is annotated with its assumability function f_j:

$$P_1^{f_1}, \ldots, P_m^{f_m} \supset Q$$

The input and output values for each f_i are nonnegative and possibly infinite. If this implication is used to backward chain from $Q_i^{c_i}$, then the literals $P_1, \ldots, P_m$ will be in the resulting formula with assumption costs $f_1(c_i), \ldots, f_m(c_i)$.

In predicate specific abduction, assumptions costs are the same for all occurrences of the predicate. Let $cost(p)$ denote the assumption cost for predicate p. The assumption cost c_i for literal Q_i in the initial formula is $cost(p)$, where the Q_i predicate is p; the assumption function f_j for literal P_j in the antecedent of an implication is the unary function whose value is uniformly $cost(p)$, where the P_j predicate is p.

In least specific abduction, different occurrences of the predicate in the initial formula may have different assumption costs, but only literals in the initial formula are assumable. The assumption cost c_i for literal Q_i in the initial formula is arbitrarily specified; the assumption function f_j for literal P_j in the antecedent of an implication has value infinity.

In chained specific abduction, the most general case, different occurrences of the predicate in the initial formula may have different assumption costs; literals obtained by backward chaining can have flexibly computed assumption costs that depend on the assumption cost of the literal backward-chained from. The assumption cost c_i for literal Q_i in the initial formula is arbitrarily specified; the assumption function f_j for literal P_j in the antecedent of an implication can be an arbitrary monotonic unary function.

We have most often used simple weighting functions of the form $f_j(c) = w_j \times c$ $(w_j > 0)$. Thus, the implication

$$P_1^{w_1} \wedge P_2^{w_2} \supset Q$$

states that P_1 and P_2 imply Q, but also that, if Q is assumable with cost c, then P_1 is assumable with cost $w_1 \times c$ and P_2 with cost $w_2 \times c$, as the result of backward chaining from Q. If $w_1 + w_2 < 1$, more specific explanations are favored, since the cost of assuming P_1 and P_2 is less than the cost of assuming Q. If $w_1 + w_2 > 1$, less specific explanations are favored: Q will be assumed in preference to P_1 and P_2. But, depending on the weights, P_i might be assumed in preference to Q if P_j is provable.

We assign to each axiom A a cost $axiom\text{-}cost(A)$ that is greater than zero. Assumption costs $assumption\text{-}cost(L)$ are computed for each literal L. When viewed abstractly, a proof is a demonstration that the goal follows from a set S of instances of the axioms, together with, in the case of abductive proofs, a set H of literals that are assumed in the proof. We want to count the cost of each separate instance of an axiom or assumption only once instead of the number of times it may appear in the syntactic form of the proof. Thus, a natural measure of the cost of the proof is

$$\sum_{A \in S} axiom\text{-}cost(A) + \sum_{L \in H} assumption\text{-}cost(L)$$

In general, the cost of a proof can be determined by extracting the sets of axiom instances S and assumptions H from the proof tree and performing the above computation. However, it is an enormous convenience if there always exists a *simple proof tree* such that each separate instance of an axiom or assumption actually occurs only once in the proof tree. That way, as the inferences are performed, costs can simply be added to

compute the cost of the current partial proof. Even if the same instance of an axiom or assumption happens to be used and counted twice, a different, cheaper derivation would use and count it only once. Partial proofs can be enumerated in order of increasing cost by employing breadth-first or iterative-deepening search methods and minimum-cost explanations can be discovered effectively.

We shall describe our inference system as an extension of pure Prolog. Prolog, though complete for Horn sets of clauses, lacks this desirable property of always being able to yield a simple proof tree.

Prolog's inference system—ordered input resolution without factoring—would have to eliminate the ordering restriction and add the factoring operation to remain a form of resolution and be able to prove Q, R from $Q \leftarrow P$, $R \leftarrow P$, and P without using P twice. Elimination of the ordering restriction is potentially very expensive.

We present a resolution-like inference system, an extension of pure Prolog, that preserves the ordering restriction and does not require repeated use of the same instances of axioms. In our extension, literals in goals can be marked with information that dictates how the literals are to be treated by the inference system, whereas in Prolog, all literals in goals are treated alike and must be proved. A literal can be marked as one of the following:

proved The literal has been proved or is in the process of being proved; in this inference system, a literal marked as proved will have been fully proved when no literal to its left remains unsolved.

assumed The literal is being assumed.

unsolved The literal is neither proved nor assumed.

The initial goal clause $Q_1, \ldots, Q_n$ in a deduction consists of literals Q_i that are either unsolved or assumed. If any assumed literals are present, they must precede the unsolved literals. Unsolved literals must be proved from the knowledge base plus any assumptions in the initial goal clause or made during the proof, or, in the case of assumable literals, may be directly assumed. Literals that are proved or assumed are retained in all successor goal clauses in the deduction and are used to eliminate matching goals. The final goal clause $P_1, \ldots, P_m$ in a deduction must consist entirely of proved or assumed literals P_i.

An abductive proof is a sequence of goal clauses $G_1, \ldots, G_p$ for which

- G_1 is the initial goal clause.
- each G_{k+1} $(1 \leq k < p)$ is derived from G_k by resolution with a fact or rule, making an assumption, or factoring with a proved or assumed literal.
- G_p has no unsolved literals.

Predicate specific abduction is quite simple because the assumability and assumption cost of a literal are determined by its predicate symbol. Least specific abduction is also comparatively simple because if a literal is not provable or assumable and must be factored, all assumable literals with which it can be factored are present in the initial and derived formulas. Because assumability is inherited in chained specific abduction, the absence of a literal to factor with is not a cause for failure. Such a literal may appear in a later derived clause after further inference as new, possibly assumable, literals are introduced by backward chaining.

Inference Rules

Suppose the current goal G_k is $Q_1^{c_1}, \ldots, Q_n^{c_n}$ and that $Q_i^{c_i}$ is the leftmost unsolved literal. Then the following inferences are possible.

Resolution with a fact

Let axiom A be a fact Q made variable-disjoint from G_k. Then, if Q_i and Q are unifiable with most general unifier σ, the goal

$$G_{k+1} = Q_1^{c_1}\sigma, \ldots, Q_n^{c_n}\sigma$$

with

$$cost'(G_{k+1}) = cost'(G_k) + axiom\text{-}cost(A)$$

can be derived, where $Q_i\sigma$ is marked as proved in G_{k+1}.

The resolution with a fact or rule operations differ from their Prolog counterparts principally in the retention of $Q_i\sigma$ (marked as proved) in the result. Its retention allows its use in future factoring.

Resolution with a rule

Let axiom A be a rule $Q \leftarrow P_1^{f_1}, \ldots, P_m^{f_m}$ made variable-disjoint from G_k. Then, if Q_i and Q are unifiable with most general unifier σ, the goal

$$G_{k+1} = \ldots, Q_{i-1}^{c_{i-1}}\sigma, P_1^{f_1(c_1)}\sigma, \ldots, P_m^{f_m(c_1)}\sigma, Q_i^{c_i}\sigma, \ldots$$

with

$$cost'(G_{k+1}) = cost'(G_k) + axiom\text{-}cost(A)$$

can be derived, where $Q_i\sigma$ is marked as proved in G_{k+1} and each $P_j\sigma$ is unsolved.

Making an assumption

The goal

$$G_{k+1} = G_k$$

with

$$cost'(G_{k+1}) = cost'(G_k)$$

can be derived, where Q_i is marked as assumed in G_{k+1}.

Factoring with a proved or assumed literal

If Q_i and Q_j $(j < i)$ are unifiable with most general unifier σ, the goal

$$G_{k+1} = \ldots, Q_j^{c_j'}\sigma, \ldots, Q_{i-1}^{c_{i-1}}\sigma, Q_{i+1}^{c_{i+1}}\sigma, \ldots$$

with

$$cost'(G_{k+1}) = cost'(G_k)$$

can be derived, where $c_j' = min(c_j, c_i)$.

Note that if Q_j is a proved literal and $c_j' < c_j$, the assumption costs of assumed literals descended from Q_j may need to be adjusted also. Thus, in resolution with a rule, it may be necessary to retain assumption costs $f_1(c_i), \ldots, f_m(c_i)$ in symbolic rather than numeric form, so that they can be readily updated if a later factoring operation changes the value of c_i.

Computing Cost of Completed Proof

If no literal of G_k is unsolved and $Q_{i_1}, \ldots, Q_{i_m}$ are the assumed literals of G_k,

$$cost(G_k) = cost'(G_k) + \sum_{i \in \{i_1, \ldots, i_m\}} c_i$$

The abductive proof is complete when all literals are either proved or assumed. Each axiom instance and assumption was used or made only once in the proof.

The proof procedure can be restricted to disallow any clause in which there are two identical proved or assumed literals. Identical literals should have been factored if neither was an ancestor of the other. Alternative proofs are also possible whenever a literal is identical to an ancestor literal.

If no literals are assumed, the procedure is a disguised form of Shostak's graph construction (GC) procedure [6] restricted to Horn clauses, where proved literals play the role of Shostak's C-literals. It also resembles Finger's ordered residue procedure [2], except that the latter retains assumed literals (rotating them to the end of the clause) but not proved literals. Thus, it includes both the ability of the GC procedure to compute simple proof trees for Horn clauses and the ability of the ordered residue procedure to make assumptions in abductive proofs.

Another approach which shares the idea of using least cost proofs to choose explanations is Post's Least Exception Logic [5]. This is restricted to the propositional calculus, with first-order problems handled by creating ground instances, because it relies upon a translation of default reasoning problems into integer linear programming problems. It finds sets of assumptions, defined by default rules, that are sufficient to prove the theorem, that are consistent with the knowledge base so far as it has been instantiated, and that have least cost.

References

[1] Cox, P.T. and T. Pietrzykowski. General diagnosis by abductive inference. *Proceedings of the 1987 Symposium on Logic Programming*, San Francisco, California, August 1987, 183–189.

[2] Finger, J.J. *Exploiting Constraints in Design Synthesis*. Ph.D. dissertation, Department of Computer Science, Stanford University, Stanford, California, February 1987.

[3] Hobbs, J.R., M. Stickel, P. Martin, and D. Edwards. Interpretation as abduction. *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, Buffalo, New York, June 1988, 95–103.

[4] Pople, H.E.,Jr. On the mechanization of abductive logic. *Proceedings of the Third International Joint Conference on Artificial Intelligence*, Stanford, California, August 1973, 147–152.

[5] Post, S.D. Default reasoning through integer linear programming. Planning Research Corporation, McLean, Virginia, 1988.

[6] Shostak, R.E. Refutation graphs. *Artificial Intelligence 7*, 1 (Spring 1976), 51–64.

[7] Stickel, M.E. Rationale and methods for abductive reasoning in natural-language interpretation. To appear in *Proceedings of the IBM Symposium on Natural Language and Logic*, Hamburg, West Germany, May 1989.

An Integrated Abductive Framework for Discourse Interpretation

Jerry R. Hobbs
Artificial Intelligence Center
SRI International

Interpretation as Abduction. Abductive inference is inference to the best explanation. The process of interpreting sentences in discourse can be viewed as the process of providing the best explanation of why the sentences would be true In the TACITUS Project at SRI, we have developed a scheme for abductive inference that yields a significant simplification in the description of such interpretation processes and a significant extension of the range of phenomena that can be captured. It has been implemented in the TACITUS System (Hobbs et al., 1988; Stickel, 1989) and has been applied to several varieties of text. The framework suggests a thoroughly integrated, nonmodular treatment of syntax, semantics, and pragmatics, and this is the focus of this paper. First, however, the use of abduction in pragmatics alone will be described.

In the abductive framework, what the interpretation of a sentence is can be described very concisely:

To interpret a sentence:

(1) Prove the logical form of the sentence,
together with the constraints that predicates impose on their arguments,
allowing for coercions,
Merging redundancies where possible,
Making assumptions where necessary.

By the first line we mean "prove from the predicate calculus axioms in the knowledge base, the logical form that has been produced by syntactic analysis and semantic translation of the sentence."

In a discourse situation, the speaker and hearer both have their sets of private beliefs, and there is a large overlapping set of mutual beliefs. An utterance stands with one foot in mutual belief and one foot in the speaker's private beliefs. It is a bid to extend the area of mutual belief to include some private beliefs of the speaker's. It is anchored referentially in mutual belief, and when we prove the logical form and the constraints, we are recognizing this referential anchor. This is the given information, the definite, the presupposed. Where it is necessary to make assumptions, the information comes from the speaker's private beliefs, and hence is the new information, the indefinite, the asserted. Merging redundancies is a way of getting a minimal, and hence a best, interpretation.

An Example. This characterization, elegant though it may be, would be of no interest if it did not lead to the solution of the discourse problems we need to have solved. A brief example will illustrate that it indeed does.

(2) The Boston office called.

This example illustrates three problems in "local pragmatics", the reference problem (What does "the Boston office" refer to?), the compound nominal interpretation problem (What is the implicit relation between Boston and the office?), and the metonymy problem (How can we coerce from the office to the person at the office who did the calling?).

Let us put these problems aside, and interpret the sentence according to characterization (1). The logical form is something like

(3) $(\exists e, x, o, b) call'(e, x) \wedge person(x) \wedge rel(x, o)$
$\wedge office(o) \wedge nn(b, o) \wedge Boston(b)$

That is, there is a calling event e by a person x related somehow (possibly by identity) to the explicit subject of the sentence o, which is an office and bears some unspecified relation nn to b which is Boston.

Suppose our knowledge base consists of the following facts: We know that there is a person John who works for O which is an office in Boston B.

(4) $person(J)$, $work\text{-}for(J, O)$, $office(O)$,
$in(O, B)$, $Boston(B)$

Suppose we also know that *work-for* is a possible coercion relation,

(5) $(\forall x, y) work\text{-}for(x, y) \supset rel(x, y)$

and that *in* is a possible implicit relation in compound nominals,

(6) $(\forall y, z) in(y, z) \supset nn(z, y)$

Then the proof of all but the first conjunct of (3) is straightforward. We thus assume $(\exists e) call'(e, J)$, and it constitutes the new information.

Notice now that all of our local pragmatics problems have been solved. "The Boston office" has been resolved to O. The implicit relation between Boston and the office has been determined to be the *in* relation. "The Boston office" has been coerced into "John, who works for the Boston office."

This is of course a simple example. More complex examples and arguments are given in Hobbs et al., 1990. A more detailed description of the method of abductive inference, particularly the system of weights and costs for choosing among possible interpretations, is given in that paper and in Stickel, 1989.

The Integrated Framework. The idea of interpretation as abduction can be combined with the older idea of parsing as deduction (Kowalski, 1980, pp. 52-53; Pereira and Warren, 1983). Consider a grammar written in Prolog style just big enough to handle sentence (2).

(7) $(\forall i, j, k) np(i, j) \wedge v(j, k) \supset s(i, k)$

(8) $(\forall i, j, k, l) det(i, j) \wedge n(j, k) \wedge n(k, l) \supset np(i, l)$

That is, if we have a noun phrase from "inter-word point" i to point j and a verb from j to k, then we have a sentence from i to k, and similarly for rule (8).

We can integrate this with our abductive framework by moving the various pieces of expression (3) into these rules for syntax, as follows:

(9) $(\forall i, j, k, e, x, y, p) np(i, j, y) \wedge v(j, k, p) \wedge p'(e, x)$
$\wedge Req(p, x) \wedge rel(x, y) \supset s(i, k, e)$

That is, if we have a noun phrase from i to j referring to y and a verb from j to k denoting predicate p, if there is an eventuality e which is the condition of p being true of some entity x (this corresponds to $call'(e, x)$ in (3)), if x satisfies the selectional requirement p imposes on its argument (this corresponds to $person(x)$), and if x is somehow related to, or coercible from, y, then there is an *interpretable* sentence from i to k describing eventuality e.

(10) $(\forall i, j, k, l) det(i, j, the) \wedge n(j, k, w_1) \wedge n(k, l, w_2)$
$\wedge w_1(z) \wedge w_2(y) \wedge nn(z, y) \supset np(i, l, y)$

That is, if there is the determiner "the" from i to j, a noun from j to k denoting predicate w_1, and another noun from k to l denoting predicate w_2, if there is a z that w_1 is true of and a y that w_2 is true of, and if there is an nn relation between z and y, then there is an *interpretable* noun phrase from i to l denoting y.

These rules incorporate the syntax in the literals like $v(j, k, p)$, the pragmatics in the literals like $p'(e, x)$, and the compositional semantics in the way the pragmatics literals are constructed out of the information provided by the syntax literals.

To parse with a grammar in the Prolog style, we prove $s(0, N)$ where N is the number of words in the sentence. To parse and interpret in the integrated framework, we prove $(\exists e) s(0, N, e)$.

Implementations of different orders of interpretation, or different sorts of interaction among syntax, compositional semantics, and local pragmatics, can then be seen as different orders of search for a proof of $(\exists e) s(0, N, e)$. In a syntax-first order of interpretation, one would try first to prove all the syntax literals, such as $np(i, j, y)$, before any of the "local pragmatic" literals, such as $p'(e, x)$. Verb-driven interpretation would first try to prove $v(j, k, p)$ and would then use the information in the requirements associated with the verb to drive the search for the arguments of the verb, by deriving $Req(p', x)$ before back-chaining on $np(i.j, y)$. But more fluid orders of interpretation are clearly possible. This formulation allows one to prove those things first which are easiest to prove, and therefore allows one to exploit the fact that the strongest clues to the meaning of a sentence can come from a variety of sources—its syntax, the semantics of its main verb, the reference of its noun phrases, and so on. The framework is, moreover, suggestive of how processing could occur in parallel, insofar as parallel Prolog is possible.

Acknowledgments. I have profited from discussions with Mark Stickel, Douglas Appelt, Stuart Shieber, Paul Martin, and Douglas Edwards about this work. The research was funded by the Defense Advanced Research Projects Agency under Office of Naval Research contract N00014-85-C-0013.

References

[1] Hobbs, Jerry R., Mark Stickel, Paul Martin, and Douglas Edwards, 1988. "Interpretation as Abduction", *Proceedings, 26th Annual Meeting of the Association for Computational Linguistics*, pp. 95-103, Buffalo, New York, June 1988.

[2] Hobbs, Jerry R., Mark Stickel, Paul Martin, and Douglas Edwards, 1990. "Interpretation as Abduction", forthcoming technical report.

[3] Kowalski, Robert, 1980. *The Logic of Problem Solving*, North Holland, New York.

[4] Pereira, Fernando C. N., and David H. D. Warren, 1983. "Parsing as Deduction", *Proceedings of the 21st Annual Meeting, Association for Computational Linguistics*, pp. 137-144. Cambridge, Massachusetts, June 1983.

[5] Stickel, Mark E. 1989. "A Prolog Technology Theorem Prover: A New Exposition and Implementation in Prolog", Technical Note No. 464. Menlo Park, Calif : SRI International.

The Role of Coherence in Constructing and Evaluating Abductive Explanations*

Hwee Tou Ng
Raymond J. Mooney

Department of Computer Sciences
University of Texas at Austin
Austin, Texas 78712
email : htng@cs.utexas.edu, mooney@cs.utexas.edu

January 1990

Abstract

Abduction is an important inference process underlying much of human intelligent activities, including text understanding, plan recognition, disease diagnosis, and physical device diagnosis. In this paper, we describe some problems encountered using abduction to understand text, and present some solutions to overcome these problems. The solutions we propose center around the use of a different criterion, called *explanatory coherence*, as the primary measure to evaluate the quality of an explanation. In addition, explanatory coherence plays an important role in the construction of explanations, both in determining the appropriate level of specificity of a preferred explanation, and in guiding the heuristic search to efficiently compute explanations of sufficiently high quality.

1 Introduction

Finding explanations for properties and events is an important aspect of intelligent behavior. The philosopher C.S. Peirce defined abduction as the process of finding the best explanation for a set of observations; i.e. inferring cause from effect. The standard formalization of abductive reasoning in artificial intelligence defines an explanation as a set of assumptions which, together with background knowledge, logically entails a set of observations [CM85].

We have built a language understanding system called ACCEL (Abductive Construction of Causal Explanations for Language) that is capable of constructing deep, causal explanations for natural language text (both narrative and expository text) through the use of abduction. ACCEL includes a generic abductive inference procedure, which computes abductive proofs by backward-chaining on the input observations using Horn-clause axioms in the knowledge base. The abductive procedure has the choice of making a subgoal in a partial proof as an assumption, if it is consistent to do so. An abductive proof represents an explanation, or an interpretation of the input sentences.

2 Problems and Solutions

2.1 Occam's Razor Isn't Sharp Enough

Almost all previous work on abduction, whether applied to plan recognition, language understanding, disease diagnosis, or physical device diagnosis, only use "Occam's Razor", i.e. the simplicity criterion, as the basis for selecting the best explanation. For instance, in [Cha86], the best interpretation is one that maxi-

*This research is supported by the NASA Ames Research Center under grant NCC-2-429. The first author was also partially supported by a University of Texas MCD fellowship. Thanks to members of the Explanation Group Meeting for helpful discussion and comments.

mizes $E - A$, where E = the number of explained observations, and A = the number of assumptions made. Other related work, though not explicitly utilizing abduction, also relies on some kind of simplicity criterion to select the best explanation. For example, [KA86] explicitly incorporates the assumption of minimizing the number of top level events in deducing the plan that an agent is pursuing.

Though an important factor, the simplicity criterion is not sufficient by itself to select the best explanation. We believe that some notion of explanatory coherence is more important in deciding which explanation is the best. This is especially true in the area of language understanding and plan recognition. In [NM89b], we have used the sentences "John was happy. The exam was easy." to illustrate this point. Relying on the simplicity metric results in selecting the interpretation that John was happy because he is an optimist, someone who always feels good about life in general (Figure 1b). This is in contrast with our preferred interpretation of the sentence — John was happy because he did well on the easy exam (Figure 1a). (See [NM89b, NM89a] for the details of the axiomatization.)

Intuitively, it seems that the first interpretation (Figure 1a) is better because the input observations are connected more "coherently" than in the second interpretation (Figure 1b). We manage to connect "John was happy" with the "easy exam" in the first interpretation, whereas in the second interpretation, they are totally unrelated. This is the intuitive notion of what we mean by *explanatory coherence*. It is clear that "Occam's Razor", i.e. making the minimum number of assumptions, is not the dominant deciding factor here at all. Rather, we select an explanation based on its coherence, i.e. how well the various observations are "tied up" together in the explanation.[1]

The notion that sentences in a natural language text are connected together in a coherent way is reflected in the well known "Grice's conversational maxims" [Gri75], which are principles governing the production of natural language utterances, such as "be

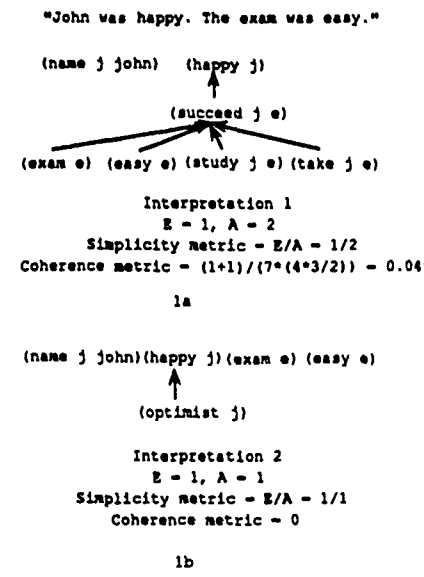


Figure 1: The importance of explanatory coherence

relevant", "be informative", etc. However, to the best of our knowledge, the work on abduction applying to the tasks of text understanding and plan recognition have not included this criterion in their evaluation of explanations. The use of explanatory coherence here attempts to remedy this problem.

We have developed a formal characterization of what we mean by explanatory coherence in the form of a coherence metric, defined as follows :

$$C = \frac{\sum_{1 \le i < j \le l} N_{i,j}}{N \binom{l}{2}}$$

where

l = the total number of observations

N = the total number of nodes in the proof graph

$\binom{l}{2} = l(l-1)/2$

$N_{i,j}$ = the number of distinct nodes n_k in the proof graph such that there is a (possibly empty) sequence of directed edges from n_k to n_i and a (possibly empty) sequence of directed edges from n_k to n_j, where n_i and n_j are observations.

We have developed and implemented an efficient algorithm to compute the coherence metric[NM89b, NM89a]. Based on the coherence metric, ACCEL has successfully selected the best interpretation for a num-

[1] Thagard [Tha89] has independently proposed a computational theory of explanatory coherence that applies to the evaluation of scientific theories. However, his theory of explanatory coherence consists of seven principles — symmetry, explanation, analogy, data priority, contradiction, general coherence, and system coherence. Independent criteria like simplicity and connectedness have been collapsed into one measure which he termed "explanatory coherence".

ber of examples of expository as well as narrative text.

2.2 Deciding on the Appropriate Level of Specificity of Explanations

Another problem in constructing a good explanation is determining the appropriate level of specificity of an abductive proof. Previous approaches fall into one of three categories : most specific abduction, least specific abduction, and weighted abduction.[2]

In most specific abduction, the assumptions made must be *basic*, i.e. they cannot be "intermediate" assumptions that are themselves provable by assuming some other (more basic) assumptions. This is the approach used in the diagnosis work of [CP87]. In least specific abduction, the only allowable assumptions are literals in the input observations. [Sti88] claims that least specific abduction is best suited for natural language interpretation. It is argued that what one learns from reading a piece of text is often close to its surface form, and that assuming deeper causes is unwarranted. In weighted abduction [HSME88], weights (or costs) are assigned to the antecedents of backward-chaining rules in order to influence the decision on whether to backchain on a rule. In this case, the best interpretation is the one with assumptions that have the lowest combined total cost.

However, none of these approaches is completely satisfactory. Consider the sentences "John went to the supermarket. He put on the uniform." Both least specific and most specific abduction fail to generate the preferred interpretation in this case, which is that John is working at the supermarket. Figure 2 shows the proof graph of the preferred interpretation of this example (excluding the dashed lines and boxes). (See [NM89a] for the details of the relevant axiomatization.)

Note that nowhere in the input sentences is the word "working" mentioned at all. It has to be *inferred* by the reader. Since this preferred interpretation includes making the assumptions that there is a working event, that John is the worker of this working event, etc, it is evident that least specific abduction, in which the only allowable assumptions are literals in the input observations, is incapable of arriving at this explanation.

[2] [Sti88] describes yet another form of abduction known as predicate specific abduction, which has been used primarily in planning and design-synthesis tasks. In predicate specific abduction, the predicate of any assumption made must be one of a pre-specified set of predicates.

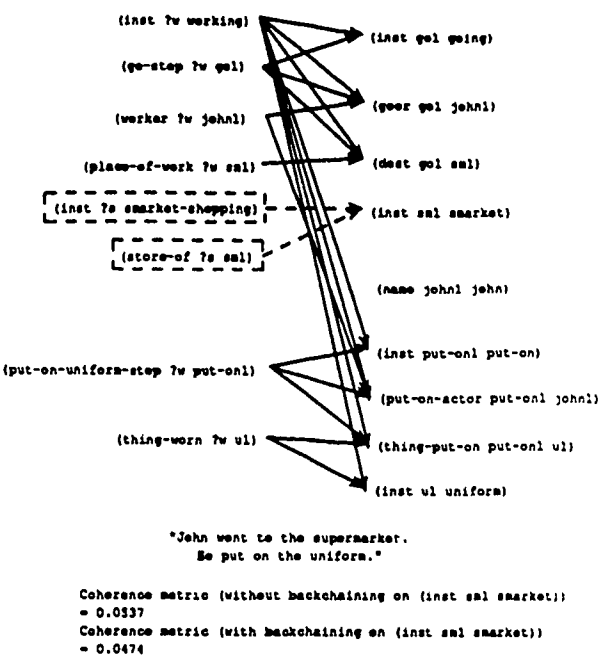


Figure 2: The level of specificity of explanation

On the other hand, most specific abduction will not do the job either. Recall that most specific abduction always prefers backchaining on rules to prove a subgoal if possible rather than making that subgoal an assumption. Thus, applying most specific abduction to this example results in backchaining on the input literal (inst sm1 smarket) to the assumptions (inst ?s smarket-shopping) and (store-of ?s sm1), since in the present knowledge base, this is the *only* backchaining rule with a consequent that unifies with (inst sm1 smarket). That is, we explain the going action, its agent and its destination by assuming that John is working there, and we are also forced to assume, by the requirement of most specific abduction, that there is some supermarket shopping event to explain the supermarket instance! This is because most specific abduction requires that we have an explanation for why John went to the supermarket as opposed to some other workplace. This is clearly undesirable.

However, determining the level of specificity of an explanation based on coherence produces the desired interpretation. That is, we backchain on rules to prove the subgoals in an explanation only if doing so increases its overall coherence, and thus we only make assumptions just specific enough to connect the ob-

servations. In the current example, backchaining on (inst sm1 smarket) results in a decrease in the coherence metric value, since the total number of nodes in the proof graph increases by two but there is no increase in the number of connections among the input observations. Intuitively, explaining the supermarket instance by assuming a supermarket shopping event is completely unrelated to the rest of the explanation that John is working there. The coherence metric has been successfully used in ACCEL to determine the appropriate level of specificity of explanations, where the desired specificity is one which maximizes coherence.

The weighted abduction of [HSME88] would presumably arrive at the correct interpretation given the "appropriate" set of weights. However, it is unclear how to characterize the "semantic contribution" of each antecedent in a rule in order to assign the appropriate weights. In contrast, our method does not rely on tweaking such weights, and it produces the preferred interpretation with the desired level of specificity in all of our examples. We believe that allowing arbitrary weights on rules is too much of a burden on the knowledge engineer. It also provides too many degrees of freedom, which can lead to the knowledge engineer "hacking up" arbitrary weights in order to get the system to produce the desired explanation.

2.3 Taming the Intractability Problem

The abduction problem has been shown to be NP-hard and so is computationally intractable [RNW85, BATJ89]. As such, the use of heuristic search to explore the vast space of possible solutions seems to be a good strategy to adopt. In fact, we have implemented a form of beam search that has successfully computed the preferred interpretation of a number of examples very efficiently.

We use a beam search algorithm which uses two beam widths, called *inter-observation beam width* (β_{inter}) and *intra-observation beam width* (β_{intra}), in order to reduce the explored search space. A queue of best explanations is kept by the beam search procedure, forming the "beam" of the beam search. At all times, explanations in the queue are sorted by coherence, where the best explanation is the one with the highest coherence.[3] Only at most β_{inter} number of the best explanations are kept in the queue after completing the processing of each input observation. Within the processing of an input observation, at most β_{intra} number of best explanations are kept in the queue.

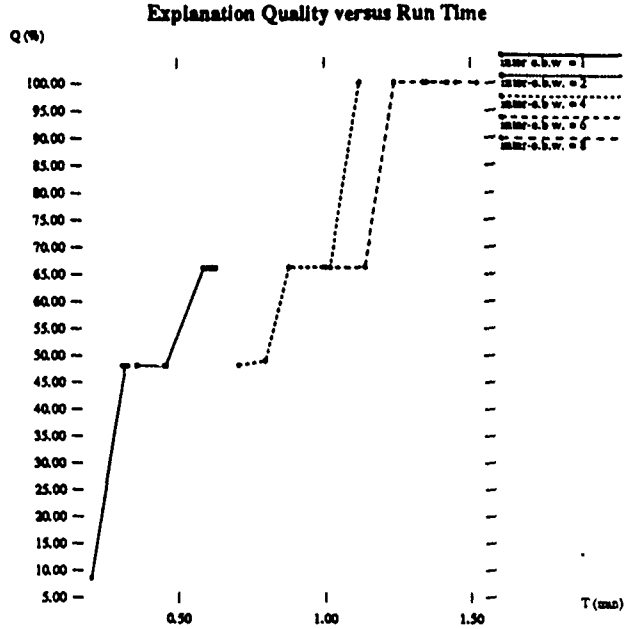


Figure 3: Explanation Quality versus Run Time

Figure 3 shows how the quality of the best explanation varies with run time for the supermarket working example by using different values of β_{inter} and β_{intra}. We use the ratio of the coherence metric value of an explanation over that of the optimal explanation to represent the quality of an explanation. All the run times reported in this paper are the actual execution times on a Texas Instruments Explorer II Lisp machine.

Each data point in the Figure represents a quality-time pair obtained by using some specific values of β_{inter} and β_{intra}. Each curve connects all the data points with the same β_{inter} but different β_{intra}. Note that without using any heuristic search (i.e. if a complete search is made), it takes more than 3 hours to compute the optimal solution, while setting $\beta_{inter} = 3$ and $\beta_{intra} = 8$ yields the optimal solution in 0.89 min. which represents a speed up of over 200 times!

[3] Ties are broken based on the simplicity metric of E/A, where E is the number of observations explained and A is the number of assumptions made.

3 Conclusion

We are looking into the possibility of making the processing more incremental by keeping track of the dependency among the assumptions and propositions of various competing explanations. Assumption-based truth maintenance systems (ATMS) [dK86] have proven useful in device diagnostic and plan recognition systems. We plan to look into the potential efficiency gain which may be brought about by incorporating an ATMS into the abductive inference procedure.

In summary, we have described some problems encountered using abduction to understand text, and have presented some solutions to overcome these problems. The solutions center around the use of explanatory coherence to evaluate the quality of explanations, to determine the appropriate level of specificity of explanations, and to guide the heuristic search to efficiently compute explanations of sufficiently high quality.

References

[BATJ89] Tom Bylander, Dean Allemang, Michael C. Tanner, and John R. Josephson. Some results concerning the computational complexity of abduction. In *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning*, pages 44–54, Toronto, Ontario, Canada, 1989.

[Cha86] Eugene Charniak. A neat theory of marker passing. In *Proceedings of the National Conference on Artificial Intelligence*, pages 584–588, Philadelphia, PA, 1986.

[CM85] Eugene Charniak and Drew McDermott. *An Introduction to Artificial Intelligence.* Addison Wesley, Reading, MA, 1985.

[CP87] P. T. Cox and T. Pietrzykowski. General diagnosis by abductive inference. In *Proceedings of the 1987 Symposium on Logic Programming*, pages 183–189, 1987.

[dK86] Johan de Kleer. An assumption-based TMS. *Artificial Intelligence*, 28:127–162, 1986.

[Gri75] H. P. Grice. Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax and Semantics 3 : Speech Acts*, pages 41–58. Academic Press, New York, 1975.

[HSME88] Jerry R. Hobbs, Mark E. Stickel, Paul Martin, and Douglas Edwards. Interpretation as abduction. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, pages 95–103, Buffalo, New York, 1988.

[KA86] Henry A. Kautz and James F. Allen. Generalized plan recognition. In *Proceedings of the National Conference on Artificial Intelligence*, pages 32–37, Philadelphia, PA, 1986.

[NM89a] Hwee Tou Ng and Raymond J. Mooney. Abductive explanation in text understanding : Some problems and solutions. Technical Report AI89-116, Artificial Intelligence Laboratory, Department of Computer Sciences, The University of Texas at Austin, October 1989.

[NM89b] Hwee Tou Ng and Raymond J. Mooney. Occam's Razor isn't sharp enough : The importance of coherence in abductive explanation. In *Proceedings of the Second AAAI Workshop on Plan Recognition*, Detroit, Michigan, August 1989.

[RNW85] James A. Reggia, Dana S. Nau, and Pearl Y. Wang. A formal model of diagnostic inference. I. problem formulation and decomposition. *Information Sciences*, 37:227–256, 1985.

[Sti88] Mark E. Stickel. A prolog-like inference system for computing minimum-cost abductive explanations in natural-language interpretation. Technical Note 451, SRI International, September 1988.

[Tha89] Paul Thagard. Explanatory coherence. *Behavioral and Brain Sciences*, 12(3):435–467, September 1989.

Problems with Abductive Language Understanding Models*

Peter Norvig and **Robert Wilensky**

University of California, Berkeley
Computer Science Division, Evans Hall
Berkeley, CA 94720, USA

Introduction

Language interpretation involves mapping from a string of words to a representation of an interpretation of those words. The problem is to be able to combine evidence from the lexicon, syntax, semantics, and pragmatics to arrive at the best of the many possible interpretations. Given the well-worn sentence "The box is in the pen," syntax may say that "pen" is a noun, while lexical knowledge may say that "pen" most often means writing implement, less often means a fenced enclosure, and very rarely means a female swan. Semantics may say that the object of "in" is often an enclosure, and pragmatics may say that the topic is hiding small boxes of illegal drugs inside aquatic birds. Thus there is evidence for multiple interpretations, and one needs some way to decide between them.

In the past few years, some general approaches to interpretation have been advanced within an abduction framework. Charniak (1986) and Norvig (1987, 1989) are two examples. In this paper we critically evaluate two later models, those of Charniak and Goldman (1989) and Hobbs, Stickel, Martin and Edwards (1988), These two models add the important property of commensurability: all types of evidence are represented in a common currency that can be compared and combined. While this is an important advance, it appears a single measure is not enough to account for all processing. We present other problems for the abductive approach, and some tentative solutions.

Cost Based Commensurability

Hobbs et al. (1988) view interpreting sentences as "providing the best explanation of why the sentences would be true." In this view a given sentence (or an entire text) is translated by an ambiguity-preserving parser into a logical form, L. Each conjunct in the logical form is annotated by a number indicating the cost, $\$C$, of assuming the conjunct to be true. Conjuncts corresponding to "new" information have a low cost of assumability, while those corresponding to "given" information have a higher cost, since to assume them is to fail to find the proper connection to mutual knowledge. Each conjunct must be either assumed or proved, using a rule or series of rules from the knowledge base. Each rule also has cost factors associated with it, and the proper interpretation, I, is the set of propositions with minimal cost that entails L.

As an example, consider again the sentence "The box is in the pen." The cost-annotated logical form (in a simplified notation omitting quantifiers) is:

$$L = box(x)^{\$10} \wedge pen(y)^{\$10} \wedge in(x,y)^{\$3}$$

where $P^{\$x}$ means the final interpretation must either assume P for $\$x$, or prove P, presumably for less. Consider the proof rules:

$$writing\,pen(x)^{\cdot 9} \supset pen(x)$$
$$enclosure(x)^{\cdot 3} \wedge fenced(x)^{\cdot 3} \wedge etc_1(x)^{\cdot 3} \supset pen(x)$$
$$female(x)^{\cdot 3} \wedge swan(x)^{\cdot 6} \supset pen(x)$$
$$enclosure(y)^{\cdot 3} \wedge inside(x,y)^{\cdot 6} \supset in(x,y)$$

The first rule says that anything that is a writing-pen is also a member of the class 'pen'—things that can be described with the word "pen". The superscripted numbers are preference information: the first rule says that $pen(x)^{\$10}$ can be derived by assuming $writing\,pen(x)^{\$9}$. Predicates of the form $etc_i(x)$, as in the second rule, denote conditions that are stated elsewhere, or, for some natural kind terms, can not be fully enumerated, but can only be assumed. They seem to be related to the abnormal predicates, $ab(x)$ used in circumscription theory (McCarthy 1986).

Below are two interpretations of L. The first just assumes the entire logical form for \$23, while the second applies the rules and shares the $enclosure(y)$ predicate common to one of the definitions of $pen(y)$ and the definition of $in(x,y)$ to arrive at a \$20.80 solution.

*Sponsored by the Defense Advanced Research Projects Agency (DoD), Arpa Order No. 4871, monitored by Space and Naval Warfare Systems Command under Contract N00039-84-C-0089. This paper benefitted from discussions with Michael Braverman, Dan Jurafsky, Nigel Ward, Dekai Wu, and other members of the BAIR seminar.

$box(x)^{\$10} \wedge pen(y)^{\$10} \wedge in(x,y)^{\$3}$.
$box(x)^{\$10} \wedge enclosure(y)^{\$3} \wedge fenced(y)^{\$3}$
$\wedge etc_1(y)^{\$3} \wedge enclosure(y)^{\$0} \wedge inside(x,y)^{\$1.8}$

The second $enclosure(y)$ gets a cost of \$0 because it has already been assumed. Let me stress that the details here are ours, and the authors may have a different treatment of this example. For example, they do not discuss lexical ambiguity, although we believe we have been faithful to the sense of their proposal.

This approach has several problems, as we see it:

(1) A single number is being used for two separate measures: the cost of the assumptions and the quality of the explanation. Hobbs et al. hint at this when they discuss the "informativeness-correctness tradeoff." Consider their example "lube-oil alarm," which gets translated as:

$$lubeoil(o)^{\$5} \wedge alarm(a)^{\$5} \wedge nn(o,a)^{\$20}$$

where nn means noun-noun compound. It is given a high cost, \$20, because failing to find the relation means failing to fully understand the referent. Intuitively this motivation is valid. However, the nn should have a very *low* cost of assumption, because there is very strong evidence for it—the juxtaposition of two nouns in the input—so there is little doubt that nn holds. Thus we see nn should have two numbers associated with it: a low cost of assumption, and a low quality of explanation. It should not be surprising to see that two numbers are needed to search for an explanation: even in A^* search one needs both a cost function, g, and a heuristic function h'.

The low quality of explanation is often the sign of a need to search for a better explanation, but the need depends on the task at hand. To diagnose a failure in the compressor, it is useful to know that a "lube-oil alarm" is an alarm that sounds when the lube-oil pressure is low, and not, say, and alarm made out of lube-oil. However, if the input was "Get me a box of lube-oil alarms from the warehouse," then it may not be necessary to further explain the nn relation.[1] Mayfield (1989) characterizes a good explanation as being applicable to the needs of the explanation's user, grounded in what is already known, and completely accounting for the input.

To put it another way, consider the situation where a magician pulls a rabbit out of his hat. One possible explanation is that the rabbit magically appeared in the hat. This explanation is of very high quality—it perfectly explains the situation—but it has a prohibitive assumption cost. An alternate explanation is that the magician somehow used slight of hand to insert the rabbit in the hat when the audience was distracted. This is of fairly low quality—it fails to completely specify the situation—but it has a much lower assumption cost. Whether this is a sufficient explanation depends on the task. For a casual observer it may will do, but for a rival magician trying to steal the trick, a better explanation is needed.

(2) Translating, say, "the pen" as $pen(y)^{\$10}$ conflates two issues: the final interpretation must find a referent, y, and it must also disambiguate "pen". It is true that definite noun phrases are often used to introduce new information, and thus must be assumed, but an interpretation that does not disambiguate "pen" is not just making an assumption—rather it is failing altogether. One could accomodate this problem by writing disambiguation rules where the sum of the left-hand-side components is less than 1. Thus, the system will always prefer to find some interpretation for "pen", rather than leaving it ambiguous. In the case of vagueness rather than ambiguity, one would probably want the left-hand-side to total greater than 1. For example, in "He saw her duck", the word "duck" is ambiguous between a water fowl and a downward movement, and any candidate solution should be forced to decide between the two meanings. In contrast, "he" is vague between a boy and a man, but it is not necessary for a valid interpretation to make this choice. We could model this with the rules:

$duck_{fowl}(x)^{\cdot 9} \supset duck(x)$
$duck_{move}(x)^{\cdot 9} \supset duck(x)$
$male(x)^{\cdot 9} \wedge adult(x)^{\cdot 2} \supset he(x)$
$male(x)^{\cdot 9} \wedge child(x)^{\cdot 2} \supset he(x)$

However, this alone is not enough. Consider the sentence "The pen is in the box." By the rules above (and assuming a box is defined as an enclosure) we could derive three interpretations, where either a writing implement, a swan, or a fenced enclosure is inside a box. All three would get a cost of \$20.8. To choose among these three, we would have to add knowledge about the likelihood of these three things being in boxes, or add knowledge about the relative frequencies of the three senses of "pen". For example, we could change the numbers as follows:

$writing\,pen(x)^{\cdot 9} \supset pen(x)$
$enclosure(x)^{\cdot 31} \wedge fenced(x)^{\cdot 31} \wedge etc_1(x)^{\cdot 31} \supset pen(x)$
$female(x)^{\cdot 4} \wedge swan(x)^{\cdot 8} \supset pen(x)$

This has the effect of making the writing implement sense slightly more likely than the fenced enclosure sense, and much more likely than the female swan sense. These rules maintain the desirable property of commensurability, but the numbers are now even more overloaded. Hobbs et al. already are giving the numbers responsibility for both "probabilities" and "semantic relatedness", and now we have shown they must account for word frequency information, and both the cost of assumptions and the quality of the explanation, the two measures needed to control search. As our previous criticisms have shown, a single number cannot represent even the cost and quality of an explanation, much less these additional factors.

[1] Translating "lube-oil alarm" as $(\exists o) lubeoil(o)$ is suspect; in the case of an alarm still in the box, there is not yet any particular oil for which it is the alarm.

Also note that to constrain search, it is important to consider bottom-up clues, as in (Charniak 1986) & Norvig 1987). It would be a mistake to use the rules given here in a strictly top-down manner, just because they are reminiscent of Prolog rules.

(3) There is no notion of a "good" or "bad" interpretation, except as an epiphenomenon of the interpretation rules. In the "pen" example, the difference between failing completely to understand "pen" and properly disambiguating it to fenced-enclosure is less than 10% of the total cost. The numbers in the rules could be changed to increase this difference, but it would still be a quantitative rather than qualitative difference. The problem is that there are at least three reasons why we might want to maintain ambiguity: because we are unsure of the cause of an event, because it is so mundane as to not need an explanation, and because it is so unbelievable that there is no explanation. This theory does not distinguish these cases. The theory has no provision for saying "I don't understand–the only interpretation I can find is a faulty one," and then looking harder for a better interpretation.

(4) There is no way to enforce a penalty worse than the cost of an assumption. Consider the sentence "Mary said she had killed herself." The logical form is something like:

$$say(Mary, x)^{\$3} \wedge x = kill(Mary, Mary)^{\$3}.$$

Thus, for $6 we can just assume the logical form, without noticing the inherent contradiction. Now let's consider some rules. We've collapsed most of the interesting parts of these rules into *etc* predicates, leaving just the parts relevant to the contradiction:

$$alive(p)^{\cdot 1} \wedge etc_2(p, x)^{\cdot 9} \supset say(p, x)$$
$$\neg alive(p)^{\cdot 5} \wedge etc_3(m, p)^{\cdot 5} \supset kill(m, p)$$

We've ignored time here, but the intent is that the alive predicate is concerned with the time interval or situation after the killing, including the time of the saying. Now, an alternative interpretation of L is:

$$alive(Mary)^{\$.3} \wedge \neg alive(Mary)^{\$1.5}$$
$$\wedge etc_2(Mary, x)^{\$2.7} \wedge etc_3(Mary, Mary)^{\$1.5}$$

Presumably there should be some penalty (finite or infinite) for deriving a contradiction, so this interpretation will total more than $6. The problem is there is no way to propagate this contradiction back up to the first interpretation, where we just assumed both clauses. We would like to penalize that interpretation, too, so that it costs more than $6, but there is no way to do so.

A solution to this problem is to legislate that rather than finding a solution to the logical form of a sentence, L, the hearer must find a solution to the larger set of propositions, L', where L' is derived from L by some process of direct, "obvious" inference. We do not want the full deductive closure from L, of course, but we want to allow for some amount of automatic forward chaining from the input.

(5) We would like to be able to go on and find alternative explanations, perhaps one where Mary is speaking from the afterworld, or she is lying, or the speaker is lying. One could imagine rules for truthful and untruthful saying, and such rules could be applied to Mary's speech act. However, since the goal of the interpretation process is "providing the best explanation of why the sentences would be true," it does not seem that we could use the rules to consider the possibility of the speaker being untruthful. The truth of the text is assumed by the model, and the speaker is not modeled.

Probability Based Commensurability

Charniak and Goldman (1988) started out with a model very similar to Hobbs et al., but became concerned with the lack of theoretical grounding for the numbers in rules, much as we were. Charniak and Goldman (1989a, 1989b) switched to a system based strictly on probabilities in the world, combined by Bayesian probability theory. Although this solves some problems, other problems remain, and some new ones are introduced. For example:

(1) The approach in (1989a) is based on "events and objects in the real world". As the authors point out, it cannot deal with texts involving modal verbs, nor can it deal with speech acts by characters, or texts where the speaker is uncooperative. So problem (4) above remains.

(2) Because the probabilities are based on events in the real world, the basic system often failed to find stories as coherent as they should be. For example, the text:

Jack got a rope. He killed himself.

suggests suicide by hanging when interpreted as a text, but when interpreted as a partial report of events in the world, that interpretation is less compelling. (After all, the killing may have taken place years after the getting.) It is only when the two events are taken as a part of a coherent text that we assume they are related, temporally and causally. In Charniak and Goldman (1989a), the coherence of stories is explained by a (probabilistic) assumption of spatio-temporal locality between events mentioned in adjacent sentences in the text. Thus the story would be treated roughly as if it were:

Jack got a rope. Soon after, nearby, a male was found to have killed himself.

The Bayesian networks compute a probability of hanging of .3; this seems about right for the later story, but too low for the original version.

Perhaps anticipating some of these problems, Charniak and Goldman (1989b) introduce an alternate approach involving a parameter, E, which denotes the probability that two arbitrary things are the same. They claim that in stories this parameter should be set higher than in real life, and that this will lead, for example, to a high probability for the interpretation where the rope that Jack got is the one he used

for hanging. But E does a poor job of capturing the notion of coherence. Consider:

John picked an integer from one to ten. Mary did so too.

Here the probability that they picked the same number should be .1, regardless of whether we are observing real life or reading a story, and regardless of the value of E.

Charniak and Goldman (1989b) go on to propose a theory of "mention" rather than a theory of coincidence, but they do not develop this alternative.

(3) It seems that for many inferences, frequency in the world does not play an important role at all. Consider the text:

Jack wanted to tie a mattress on top of his car. He also felt like killing himself. He got some rope.

Now, the probability of getting a rope to hang oneself given suicidal feelings must be quite low, maybe .001, while the probability of getting a rope for tying given a desire to secure a mattress is much higher, maybe .5. Thus the Charniak-Goldman model would strongly prefer the latter interpretation. With the "mention" theory, it would like both interpretations. Yet a sample of informants mostly found the text confusing–they reported finding both interpretations, and were unable to choose between them. It would be useful to find a better characterization of when frequencies in the world are useful, and when they appear to be ignored in favor of some more discrete notion of "reasonable connection."

Problems With Both Models

Neither model is completely explicit on how the final explanation is constructed, or on what to do with the final explanation. In a sense, Hobbs et al.'s system is like a justification-based truth-maintenance system that searches for a single consistent state, possibly exploring other higher-cost states along the way. Charniak and Goldman's system is like an assumption-based truth-maintenance system (ATMS) that keeps track of all possible worlds in one grand model, but needs a separate interpretation process to extract consistent solutions. Thus, the system does not really do interpretation to the level that could lead to decisions. Rather, it provides evidence upon which decisions can be based.

Both approaches are problematic. Imagine the situation where a hearer is driving a car, and is about to enter an intersection when a traffic officer says "don't - stop". The hearer derives two possible interpretations, one corresponding to "Don't stop." and the other corresponding to "Don't. Stop." Hobbs et al.'s system would assign costs and chose the one with the lower cost, no matter how slight the difference. A more prudent course of action might be to recognize the ambiguity, and seek more information to decide what was intended. Charniak and Goldman's system would assign probabilities to each proposition, but would offer no assistance as to what to do. However, if the model were extended from Bayesian networks to influence diagrams, then a decision could be made, and it would also be possible to direct search to the important parts of the network.

Deliberate ambiguity is also a problematic area. In a pun, for example, the speaker intends that the hearer recover two distinct interpretations. Such subtlety would be lost on the models discussed here. This issue is discussed in more depth in Norvig (1988).

A number of arguments show that strict maximization of probability (or minimization of cost) is a bad idea. First, as we have seen, we must sometimes admit that an input is truly ambiguous (intentionally or unintentionally).

Second, there is the problem of computational complexity. Algorithms that guarantee a maximal solution take exponential time for the models discussed here. Thus, a large-scale system will be forced to make some sort of approximation, using a less costly algorithm. This is particularly true because we desire an on-line system—one that computes a partial solution after each word is read, and updates the solution in a bounded period of time.

Third, communication by language has the property that "the speaker is always right". In chess, if I play optimally and my opponent plays sub-optimally, I win. But in language understanding, if I abduce the "optimal" interpretation when the speaker had something else in mind, then we have failed to communicate, and I in effect lose. Put another way, there is a clear "evolutionary" advantage for optimal chess strategies, but once language has evolved to the point where communication is possible, there is no point for a hearer to try to change his interpretation strategy to derive what an optimal speaker would have uttered to an optimal hearer–because there are no such optimal speakers. Indeed, there is an advantage for communication strategies that can be computed quickly, allowing the participants to spend time on other tasks. By the second point above, such a strategy must be sub-optimal.

Earlier we said that Charniak and Goldman (1989b) introduced the parameter E to account for the coherence of stories. But they also provide a brief sketch of another account, one where, in addition to deriving probabilities of events in the world, we also consider the probability that the speaker would mention a particular entity at all. Such a theory, if worked out, could account for the difficulty in processing speech acts that we have shown both models suffer from.

However, a theory of "mention" alone is not enough. We also need theories of representing, intending, believing, directly implying, predicting, and acting. The chain of reasoning and acting includes at least the following:

H attends to utterance U by speaker S

H infers "S said U to H"

H infers "L represents U"

H infers "L directly implies L'"
H infers "S intended H to believe S believes L'"
H infers "S intended H to believe L'"
H believes a portion of L' compatible with H's beliefs
H forms predictions about S's future speech acts
H acts accordingly

This still only covers the case of successful, cooperative communication, and it leaves out some steps. A successful model should be able to deal with all these rules, when necessary. However, the successful model should also be able to quickly bypass the rules in the default case. We believe that the coherence of stories stems primarily from the speaker presenting evidence to the hearer in a fashion that will lead the hearer to focus his attention on the evidence, and thereby derive the inferences intended by the speaker. Communication is possible because it consists primarily of building a single shared explanation. It is only in unusual cases where there are multiple possibilities that must be weighed against each other and carried forth.

Both models seem to have difficulty distinguishing ambiguity from multiple explanations. This makes a difference in cases like the following:

John was wondering about lunch when it started to rain. He ran into a restaurant.

Here there are two reasons why John would enter the restaurant–to satisfy hunger and to avoid the rain. In other words there are two explanations, say, $A \supset R$ and $B \supset R$, and we would like to combine them to yield $A \wedge B \supset R$. As we understand it, Hobbs et al. appear to use "exclusive or" in all cases, so they would not find this explanation. Charniak and Goldman allow competing explanations to be joined by an "or" node, but require competing lexical senses to be joined by "exclusive or" nodes. So they would find $A \vee B \supset R$. In other words, they would find both explanations probable, which is not quite the same thing as finding the conjunction probable. Now consider:

He's a real sweetheart.

This has a straight and an ironic reading: $sweetheart(x)$ and $\neg sweetheart(x)$. The disjunction is a tautology and the conjunction is a contradiction, so in this case the Hobbs approach of keeping the alternatives separate seems better than allowing their disjunction. Finally, consider:

Mary was herding water fowl while dodging hostile gunfire. John saw her duck.

Here we do not want to combine two the interpretations into a single interpretation. If we amend a model to allow multiple explanations, we must be careful that we don't go too far.

Conclusions

Abduction is a good model for language interpretation, and commensurability is a vital component of an abduction system. But the models discussed here have serious limitations, due to technical problems, and due to a failure to embrace language as a complex activity, involving actions, goals, beliefs, inferences, predictions, and the like. We don't believe that knowledge of probability in the world, plus a few general principles (such as E) can lead to a viable theory of language use. This "complicated" side of language has been studied in depth for over a decade (a list very similar to our chain of reasoning and acting appears in Morgan (1978)), so our task is clear: to marry these pre-theoretic "complicated" notions with the formal apparatus of commensurable abductive interpretation schemes.

References

Charniak, E. A neat theory of marker passing, *AAAI-86.*

Charniak, E. and Goldman, R. (1988) A logic for semantic interpretation, *Proc. of the 26th Meeting of the ACL.*

Charniak, E. and Goldman, R. (1989a) A semantics for probabilistic quantifier-free first-order languages, with particular application to story understanding, *IJCAI-89.*

Charniak, E. and Goldman, R. (1989b) Plan recognition in stories and in life, *Uncertainty Workshop, IJCAI-89.*

Hobbs, J. R., Stickel, M., Martin, P. and Edwards, D. (1988) Interpretation as abduction, *Proc. of the 26th Meeting of the ACL.*

Mayfield, J. M. (1989) Goal analysis: Plan recognition in dialogue systems, *Univ. of Cal. Berkeley EECS Dept.* Report No. UCB/CSD 89/521.

McCarthy, J. (1986) Applications of circumscription to formalizing common-sense knowledge. *Artificial Intelligence,* 26(3).

Morgan, J. L. (1978) Toward a rational model of discourse comprehension. *Theoretical Issues in Natural Language Processing.*

Norvig, P. (1987) A Unified Theory of Inference for Text Understanding. *Univ. of Cal. Berkeley EECS Dept.* Report No. UCB/CSD 87/339.

Norvig, P. (1988) Multiple simultaneous interpretations of ambiguous sentences. *Proc. of the 10th Annual Conference of the Cognitive Science Society.*

Norvig, P. (1989) Marker passing as a Weak Method for Text Inferencing. *Cognitive Science,* 13, 4, 569-620.

Abductive Speech Act Recognition

Elizabeth Ann Hinkelman
Center for Information and Language Studies
University of Chicago, 1100 East 57th Street,
Chicago, Illinois 60637

Introduction

My recent dissertation [Hinkelman, 1989] describes an application of automated abduction to speech act recognition. It includes a unification pattern matching component, which allows lexical and syntactic cues to suggest possible speech act interpretations, and a weighted heuristic search component which explores an inference space of plan recognition rules. Much of the interest of this system is that it can handle a broad range of examples precisely because of the integration of two abduction techniques. My experiences with this system have led to a number of conclusions about automated abduction.

Speech Act Recognition

The problem of reconstructing an agent's intentional structure from observed actions, or plan recognition, is a fundamental application of automated abduction. Plan recognition occurs in the domain of natural language processing in two forms. The more obvious is the reconstruction of plans and goals that are unrelated to language, from linguistic observations. I will refer to such goals and plans as domain plans. Domain plans may be described in texts such as stories, or discussed as in ordinary talk. In ordinary talk, however, it is necessary to know whether a described action is being suggested, requested, asserted, denied, forbidden, and so on. Thus the second type of plan recognition arises. It becomes useful to view the utterances themselves as actions, from which communicative intentions can be recognized. My work to date has concentrated on the recognition of such speech acts [Searle, 1969], and integrates general, theoretically powerful mechanisms with a more specialized linguistic scheme for broad coverage of the phenomena.

The input utterances are first processed by the purely linguistic scheme, which makes no use of context. The linguistic scheme consists of unification-based pattern matching. Patterns of linguistic features are matched against a pre-parsed string, yielding sets of partial speech act descriptions. The descriptions are composed using unification, producing a set of possible speech acts. The linguistic rules allow for the arbitrary nature of linguistic conventions, but treat these conventions uniformly as incremental evidence for some range of actions. This method allows the system to distinguish between "Can you pass me the salt?", which is likely a request, and "Are you able to pass me the salt?" which is likely a yes/no question.

The more general method is based on Allen's [Allen, 1983]. It takes an action as input and uses weighted heuristic search through a space of plan recognition inferences. The inference rules are obtained by inverting a set of plan construction rules. This general method is appropriate for domain plan recognition as well as speech act recognition, and serves as a backup to the more efficient, specialized linguistic scheme. It captures the relationship between asserting "I need X" and requesting the hearer to provide X, making full use of contextual information.

Integration between these two components is achieved by invoking the general method only when the linguistic method fails to provide a suitable interpretation. 'Suitability' is determined by a mechanism interesting in its own right; please see the accompanying papers for details. The above is simply background for subsequent discussion points.

Discussion

Experience with abductive speech act recognition leads to several observations.

A Question of Methodology

The goal of this work is to provide a model of human language processing. It addresses the phenomenon of "indirect" speech acts, and as such must provide an account of various classes of these acts and how they are identified from what is said. I therefore evaluate the system according to its ability to provide the same identifications of speech acts that people do. Presently "what people do" is as much a matter of the linguist's intuitions as of psycholinguistic studies of subjects [Gibbs,

1986] and one hopes that more psycholinguistic studies will appear. With appropriate architectural considerations, tasks which are easy for people should be easy for the system, and tasks which are difficult for people should be difficult for the system. Such evaluation methods are clearly inappropriate for systems which seek to improve on human performance.

The Two-Level Architecture

Two-level systems such as this one are instances of a general principle of computer systems design [Lampson, 1983], which specifies that the vast majority of ordinary tasks should be performed by an ordinary method which is kept simple and efficient by offloading the difficult cases to a more powerful, expensive, and rarely invoked mechanism. Care must be taken to allow the two components to integrate smoothly. For those abduction applications in which there is a similar division of tasks into common, simple and uncommon, difficult tasks, a two-level design would be appropriate.

In an abductive planning system, for example, the output plans can be viewed as an "explanation" of environmental stimuli in terms of their role in meeting the agent's goals. One can imagine a planning system that has a reactive component, a production system mapping sensory stimuli onto rather immediate actions, and a plan construction component, which may screen immediate actions or search for longer-term methods. Such an approach is being pursued by Feist [Feist,].

Knowledge Representation

The two-level design concept applies to the underlying knowledge representation as well. In the speech act interpreter, for example, the pattern matching component may be amenable to a stochastic, massively parallel treatment such as a connectionist network. Stochastic massive parallelism shows great promise as a form of knowledge representation suitable for artificial intelligence. However, as a methodological strategy discrete methods avoid certain tedious knowledge engineering tasks and promote clear, testable theories. They also avoid unresolved issues in connectionist representations, such as the role and method of variable binding, or the incorporation of the temporal continuity of input. For example, most connectionist models of word recognition must replicate their structure for subsequent time intervals [McClelland and Elman, 1986]. It is interesting to note that FOPC has the inverse problem with respect to variable binding, a solution to which was proposed by Charniak [Charniak,].

The plan reasoning component requires representation of very general inference patterns which are much more difficult to formulate in such low-level terms. Advances in knowledge representation may someday illuminate a relationship between a slow, serial reasoner and the 'lower' level, and this would make it much easier to explain how learning paths of reasoning could lead to learning of lower level correlations, and vice versa. My theory of speech act recognition includes the claim that many of the correlations embodied in the linguistic component do in fact originate from extended inference; one way of doing this was described by Pazzani in the explanation-based learning paradigm [Pazzani,]. Such learning may not be desireable in all abductive inference domains, but in those which model intelligent agents the flexibility is crucial. Serial, inference-based methods may be slow or brittle, but the main weakness of the speech act inference component is in the area of controlling search.

Inference Methods

The relationships among methods of inference for plan recognition have been discussed in detail by Kautz [Kautz, 1987]. Kautz himself presents a deductive method of plan recognition, with circumscription. The data structure is a hierarchy of (multistep) actions, defined by an abstraction (is-a) and decomposition (step of) relation. For each observed action, the relations are used to identify all possible ways for this action to participate in actions marked as being ends in themselves. A series of observed actions can be explained as the minimal set of end actions necessary to account for these steps. The best explanation is defined as the most parsimonious, corresponding to minimization of the End predicate in the model theory. Further criteria for a best explanation are given, but without correspondence to a feature of the model theory.

Kautz shows how this circumscription method can be applied to domain plan recognition in the cooking domain, to medical diagnosis, and to speech act recognition. I cannot speak to the aptness of the medical diagnosis application. For speech act recognition, what I have found is that speech acts have a role in ordinary domain plans which is ad hoc rather than prototypical to these plans. This makes incorporation of speech act steps expensive because they would have to be inserted at every point where they may rarely be needed. And although it may be possible to construe the linguistic features as action observations, using linguistic pattern matching rather than the step-of relation provides more robustness in accounting for the variable phenomena. Thus we sacrifice the model-theoretic semantics, which in any case did not incorporate all aspects of a best explanation.

Kautz's method has a clearly specified inference procedure. It therefore has better-defined results than dynamic logic or default logic, in which the results depend on the order of rule application. Statistical methods require an acceptance rule, such as "accept the explanation with the highest probability", and counterexamples to any such rule seem inevitable in real applications. Thus although all of these methods show promise, they all have remaining difficulties.

A Problem of Belief Revision

A fundamental feature of abductive methods is that what appears to be the best explanation of some phenomenon may later prove to be wrong. Subsequent input can therefore require non-incremental changes to the state of knowledge, if explanations are incorporated into stored knowledge. Methods of truth maintenance have been proposed to allow retraction of explanations if any of their links are invalidated. But even when applicable, these methods too leave unspecified how to determine which explanation is "better". The problems in detecting a need for belief revision, arbitrating it, and updating the database can all be formidable.

Conclusion

Application of abduction to speech act recognition has used a methodology oriented toward obtaining performance analogous to that of human intuitions and behavior. It has shown the utility of a two-level system in which the common cases are handled efficiently and the difficult ones with greater power. It raises specific problems for current knowledge representation and inference methods. And it may yet be illuminated by reports on abductive methods from other areas.

References

[Allen, 1983] James F. Allen. Recognizing intentions from natural language utterances. In M. Brady and R. C. Berwick, editors, *Computational Models of Discourse*. MIT Press, Cambridge, MA, 1983.

[Charniak,] Eugene Charniak. Motivation analysis, abductive unification, and non-monotonic equality. Unpublished Manuscript, Brown University, Providence, RI.

[Feist,] Steven Feist. Integrating symbolic planning and reactive execution. Thesis Proposal, University of Rochester Department of Computer Science, 1989.

[Gibbs, 1986] Ray W. Gibbs. What makes some indirect speech acts conventional? *Journal of Memory and Language*, 15:181–196, 1986.

[Hinkelman, 1989] Elizabeth Ann Hinkelman. Linguistic and pragmatic constraints on utterance interpretation. Technical Report TR 288, University of Rochester Department of Computer Science, 1989.

[Kautz, 1987] Henry A. Kautz. A formal theory of plan recognition. Technical Report TR 215, University of Rochester Department of Computer Science, May 1987.

[Lampson, 1983] Butler W. Lampson. Hints for computer systems design. *Operating Systems Review*, 17(5):33–48, 1983.

[McClelland and Elman, 1986] J. L. McClelland and J. L. Elman. Interactive processes in speech perception: The trace model. In J.L. McClelland and D. E. Rumelhart, editors, *Parallel Distributed Processing*. MIT Press, Cambridge, MA, 1986.

[Pazzani,] M. Pazzani. Learning indirect speech acts. Submitted to Association for Computational Linguistics, 1989.

[Searle, 1969] John Searle. *Speech Acts*. Cambridge University Press, New York, 1969.

Goal-based explanation

Ashwin Ram

Georgia Institute of Technology
School of Information and Computer Science
Atlanta, Georgia 30332-0280
(404) 853-9372
ashwin@gatech.edu

In order to learn from experience, a reasoner must be able to *explain* what it does not understand. When a novel or poorly understood situation is processed, it is interpreted in terms of knowledge structures already in memory. As long as these structures provide expectations that allow the reasoner to function effectively in the new situation, there is no problem. However, if these expectations fail, the reasoner is faced with an *anomaly*. The world is different from its expectations. In order to learn from this experience, the reasoner needs to know *why* it made those predictions. It also needs to explain *why* the failure occurred, i.e., to identify the knowledge structures that gave rise to the faulty expectations, and to understand why its domain model was violated in this situation. Finally, it must store the new experience in memory for future use. *Abduction*, the construction of explanations, is a central component of this learning process.

Abduction is often viewed as inference to the "best" explanation. However, the definition of "best" is dependent on the *goals* of the reasoner in forming the explanation and not just on the correctness of the causal chain underlying the explanation. In situations where there is no one "right" explanation, the "best" explanation must be more than a causal chain that describes the domain; it must also address the reason that an explanation was required in the first place. This in turn determines what the reasoner can learn from the explanation.

What is an explanation?

The need for an explanation arises when some observed fact doesn't quite fit into the reasoner's world model, i.e., the reasoner detects an *anomaly*. An explanation is a knowledge structure that makes the anomaly go away. To illustrate the nature of such a structure, let us consider some candidate explanations for the anomaly underlying the following popular joke:

S-1: Why do firemen wear red suspenders?

(1) Because it is always raining in New Haven.

(2) To keep their pants up.

(3) Because red, the symbol of warning, is the color of the fire brigade's uniform.

(4) Because red suspenders look funny if they aren't part of a uniform.

Consider (1). This does not seem like an explanation for S-1. The reason isn't that (1) is false, but rather that there seems to be no causal connection between (1) and S-1. Thus it is not sufficient for a proposed explanation to be true; *an explanation must be causally connected to the anomaly.* It must contain a set of premises and a causal chain linking those premises to the anomalous proposition. If the reasoner believes the premises, the proposition ceases to be anomalous since the causal interactions underlying the situation can now be understood.

However, not all causal structures are explanations. For example, (2) is causally relevant to S-1, but it still doesn't feel like an explanation. To understand why, let us make the anomaly in S-1 explicit. The real question isn't "Why do firemen wear red suspenders?", but rather one of the following:

S-2: Why do firemen wear only red suspenders? If firemen are a representative sample of the general population, we would expect them to wear suspenders of various colors, and even belts.

S-3: Why doesn't everyone wear red suspenders? If red suspenders are indeed attractive or desirable, we would expect everyone to wear red suspenders, not just firemen.

The reason that the joke is funny is that (2) misses the point of the question. If the point is made explicit as in S-2, (3) is a possible explanation for the anomaly. Alternatively, if the real question is intended to be S-3, (4) is a possible explanation. The point is that, in order to qualify as an explanation, a causal description must address the underlying anomaly.

To state this another way, *an explanation must address the failure of the reasoner to model the situation correctly.* In addition to resolving the incorrect predictions, it must also point to the erroneous aspect of

the chain of reasoning that led to the incorrect predictions. An explanation is *useful* if it allows the reasoner to learn;[1] the claim here is that *an explanation must be both causal and relevant in order to be useful.*

An explanation, therefore, must address two types of questions:

1. Why did things occur as they did in the world? This question gives rise to *knowledge acquisition goals*, which are goals to collect information or knowledge about the domain that the anomaly has signalled as being missing.
2. Why did I fail to predict this correctly? This question gives rise to *knowledge organization goals*, which are goals to improve the organization of knowledge in memory.

The answer to the first question is called a *domain explanation* since it is a statement about the causality of the domain. The answer to the second question is called an *introspective* or *meta-explanation* since it is a statement about the reasoning processes of the system. The claim here is that an explanation must supply both answers in order to be useful. Let us consider the second one first.

Introspective explanations: Addressing knowledge organization goals

One of the questions an explanation must address is why the reasoner failed to make the correct prediction in a particular situation. This could happen in one of the following ways:

1. **Novel situation:** The reasoner did not have the knowledge structures to deal with the situation.
2. **Incorrect world model:** The knowledge structures that the reasoner applied to the situation were incomplete or incorrect.
3. **Mis-indexed domain knowledge:** The reasoner did have the knowledge structures to deal with the situation, but it was unable to retrieve them since they were not indexed under the cues that the situation provided.

When an explanation is built, the reasoner needs to be able to identify the kind of processing error that occurred and invoke the appropriate learning strategy. For example, if an incomplete knowledge structure is applied to a situation, the resulting processing error represents both the knowledge that is missing, as well as the fact that this piece of knowledge, when it comes in, should be used to fill in the gap in the original knowledge structure. Similarly, if an error arose due to a mis-indexed knowledge structure, the explanation, when available, should be used to re-index the knowledge structure appropriately.

Knowledge organization goals can be categorized by the type of gap that gave rise to them, or by the type of learning that results from their satisfaction:

1. **Missing knowledge** – learn new knowledge to fill gap in domain model
2. **Unconnected knowledge** – learn new connection or new index
3. **Implicit assumption** – learn heuristics for when to check assumption explicitly
4. **Calculated simplification** – learn heuristics for when to check assumption in detail
5. **Explicit assumption** – learn new knowledge to correct the assumption
6. **Conjunctive assumptions** – learn new interactions

Domain explanations: Addressing knowledge acquisition goals

Knowledge acquisition goals seek causal knowledge about the domain. A domain explanation is a causal chain that demonstrates why the anomalous proposition might have occurred by introducing a set of premises that causally lead up to that proposition. If the reasoner believes or can verify the premises of an explanation, the conclusion is said to be explained. Explanations are often verbalized using their premises. However, the real explanation includes the premises, the causal chain, and any intermediate assertions that are part of the causal chain.

Domain explanations can be divided into two broad categories, physical and volitional.

Physical explanations Physical explanations link events with the states that result from them, and further events that they enable, using causal chains similar to those of [Rieger, 1975] and [Schank and Abelson, 1977]. Physical explanations answer questions about the physical causality of the domain.

Volitional explanations Volitional explanations link actions that people perform to their goals and beliefs, yielding an understanding of the *motivations* of the characters. Volitional explanations thus correspond to the filling out of the "belief-goal-plan-action" chain [Schank and Abelson, 1977; Wilks, 1977; Wilensky, 1978; Schank, 1986], although we need to expand the vocabulary of this chain in order to model such explanations adequately [Ram, 1989]. A volitional explanation relates the actions in which the characters in the story are involved to the *outcomes* that those actions had for them, the *goals*, *beliefs*, *emotional states* and *social states* of the characters as well as priorities or *orderings* among the goals, and the *decision process* that the characters go through in *considering* their goals,

[1] Learning is often performed in the service of a problem-solving task; thus knowledge goals of the type described here often arise from the problem-solving goals of the reasoner. This issue is beyond the scope of this paper.

goal-orderings and likely outcomes of the actions before deciding whether to do those actions. A detailed volitional explanation involving the planning decisions of a character is called a *decision model* [Ram, 1989].

Decision models provide a theory of motivational coherence for stories involving volitional agents. When a decision model is applied to the actions of a given character in a story, it may give rise to questions based on faulty assumptions or inconsistencies identified in the application of the decision model to the story. These inconsistencies signal anomalies, which must be explained by determining whether different parts of the decision model (e.g., the goals of the agent, his beliefs about the outcome, or his volition in deciding to perform the action) are actually present as assumed. These anomalies give rise to a set of knowledge acquisition goals which the reasoner tries to satisfy by building volitional explanations.

Components of explanation patterns

Standard domain explanations known to the reasoner are called *explanation patterns* [Schank, 1986]. Explanation patterns (XPs) have four main components [Ram, 1989]:

1. **PRE-XP-NODES:** Nodes that represent what is known before the XP is applied. One of these nodes, the EXPLAINS node, represents the particular action being explained.
2. **XP-ASSERTED-NODES:** Nodes asserted by the XP as the explanation for the EXPLAINS node. These comprise the premises of the explanation.
3. **INTERNAL-XP-NODES:** Internal nodes asserted by the XP in order to link the XP-ASSERTED-NODES to the EXPLAINS node.
4. **LINKS:** Causal links asserted by the XP. These taken together with the INTERNAL-XP-NODES are also called the internals of the XP.

An explanation pattern states that the XP-ASSERTED-NODES lead to the EXPLAINS node (which is part of a particular configuration of PRE-XP-NODES) via a set of INTERNAL-XP-NODES, the nodes being causally linked together via the LINKS. In other words, an XP is a causal chain composed of a set of nodes connected together using a set of LINKS (causal rules or XPs). The "antecedent" of this causal chain is the set of XP-ASSERTED-NODES, the "internal nodes" of the causal chain are the INTERNAL-XP-NODES of the XP, and the "consequent" is the EXPLAINS node. The difference between XP-ASSERTED-NODES and INTERNAL-XP-NODES is that the former are merely asserted by the XP without further explanation, whereas the latter have causal antecedents within the XP itself.

The explanation cycle

An explanation-based understander must be able to detect anomalies in the input, and resolve them by building motivational and causal explanations for the events in the story in order to understand why the characters acted as they did, or why certain events occurred or did not occur. This process characterizes both "story understanders" that try to achieve a deep understanding of the stories that they read, as well as programs that need to understand their domains in service of other problem-solving tasks.

The process model for the task of explanation consists of the following steps:

1. **Anomaly detection:** Anomaly detection refers to the process of identifying an unusual fact that needs explanation. The anomalous fact may be unusual in the sense that it violates or contradicts some piece of information in memory. Alternatively, the fact may be unusual because, while there is no explicit contradiction, the reasoner fails to integrate the fact satisfactorily in its memory.
2. **Explanation pattern retrieval:** When faced with an anomalous situation, the reasoner tries to retrieve one or more explanation patterns that would explain the situation. These patterns could be abstract causality templates, such as those of [Schank, 1986], or descriptions of causality underlying specific cases known to the reasoner, such as those used by case-based reasoners (e.g., [Kolodner, 1988; Hammond, 1989]).
3. **Explanation pattern application:** Once a set of potentially applicable explanation patterns is retrieved, the reasoner tries to use them to resolve the anomaly. This involves instantiating the XP, filling in the details through elaboration and specification, and checking the validity of the final explanation. An XP is instantiated by unifying the EXPLAINS node of the XP with the description of the situation being explained, and instantiating the INTERNAL-XP-NODES and LINKS. If all the PRE-XP-NODES and INTERNAL-XP-NODES of the XP fit the situation, the hypothesis is applicable. If the unification fails, the hypothesis is rejected.[2]
4. **Hypothesis verification:** The final step in the explanation process is the confirmation or refutation of possible explanations, or, if there is more than one hypothesis, discrimination between the alternatives. A hypothesis is a causal graph that connects the premises of the explanation to the conclusions via a set of intermediate assertions. At the end of this step, the reasoner is left with one or more alternative hypotheses. Partially confirmed hypotheses are maintained in a data dependency network called a

[2] There is also the possibility of modifying the hypothesis to fit the situation [Schank, 1986; Kass *et al.*, 1986].

hypothesis tree, along with questions (knowledge acquisition goals) representing what is required to verify these hypotheses.

Evaluating explanations

There are five criteria for evaluating the goodness of an explanation:

1. **Believability:** Do I believe the XP from which the hypothesis was derived? This is not an issue when all XPs in memory are believed, but for a program that learns new XPs, some of which may be incomplete, the believability of the XP is an important criterion in deciding whether to believe the resulting hypothesis.
2. **Applicability:** How well does the XP apply to this situation? Did it fit the situation without any modifications?
3. **Relevance:** Does the XP address the underlying anomaly? Does it address the knowledge goals of the reasoner?
4. **Verification:** How definitely was the explanation confirmed or refuted?
5. **Specificity:** How specific is the XP? Is it abstract and very general (e.g., a proverb), or is it detailed and specific?

Intuitively, a "good" explanation is not necessarily one that can be proven to be "true" (criterion 4), but also one that seems plausible (1 and 2), fits the situation well (2 and 5), and is relevant to the goals of the reasoner (criterion 3).

Conclusion

Abduction, or inference to the best explanation, is a central component of the reasoning process. The "best" explanation is not one that is the most "correct," if correctness is even measurable in the domain of interest, but one that is most useful to the process that is seeking the explanation.

These ideas have been explored in the AQUA program, a computer model of the theory of question-driven understanding [Ram, 1989; Ram, 1987; Schank and Ram, 1988]. AQUA learns about terrorism by reading newspaper stories about terrorist incidents in the Middle East. AQUA's model of terrorism is never quite complete; knowledge structures may have "gaps" in them, or they may not be indexed correctly in memory. When AQUA reads a story, these gaps give rise to questions about the input. The point of reading is find answers to these questions, to learn by filling in the gaps in its world model.

Questions, therefore, represent the "knowledge goals" of the understander, things that the understander wants to learn about. AQUA builds explanations in order to find answers to its questions. Thus AQUA is an example of a system based on the goal-directed explanation process presented in this paper.

References

[Hammond, 1989] Kristian J. Hammond, editor. *Proceedings of a Workshop on Case-Based Reasoning.* Morgan Kaufmann, Inc., Pensacola Beach, FL, May 1989.

[Kass *et al.*, 1986] Alex Kass, David Leake, and Christopher Owens. *SWALE: A Program That Explains*, pages 232–254. Lawrence Erlbaum Associates, Hillsdale, NJ, 1986.

[Kolodner, 1988] Janet L. Kolodner, editor. *Proceedings of a Workshop on Case-Based Reasoning.* Morgan Kaufmann, Inc., Clearwater Beach, FL, May 1988.

[Ram, 1987] Ashwin Ram. Aqua: Asking questions and understanding answers. In *Proceedings of the Sixth Annual National Conference on Artificial Intelligence*, pages 312–316, Seattle, WA, July 1987. American Association for Artificial Intelligence, Morgan Kaufman Publishers, Inc.

[Ram, 1989] Ashwin Ram. *Question-driven understanding: An integrated theory of story understanding, memory and learning.* PhD thesis, Yale University, New Haven, CT, May 1989. Research Report #710.

[Rieger, 1975] C. Rieger. Conceptual memory and inference. In Roger C. Schank, editor, *Conceptual Information Processing.* North-Holland, Amsterdam, 1975.

[Schank and Abelson, 1977] Roger C. Schank and Robert Abelson. *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures.* Lawrence Erlbaum Associates, Hillsdale, NJ, 1977.

[Schank and Ram, 1988] Roger C. Schank and Ashwin Ram. Question-driven parsing: A new approach to natural language understanding. *Journal of Japanese Society for Artificial Intelligence*, 3(3):260–270, May 1988.

[Schank, 1986] Roger C. Schank. *Explanation Patterns: Understanding Mechanically and Creatively.* Lawrence Erlbaum Associates, Hillsdale, NJ, 1986.

[Wilensky, 1978] Robert Wilensky. *Understanding Goal-Based Stories.* PhD thesis, Yale University, Department of Computer Science, New Haven, CT, 1978.

[Wilks, 1977] Yorick Wilks. What sort of taxonomy of causation do we need for language understanding. *Cognitive Science*, 1:235, 1977.

Integrating Abduction and Learning

Paul O'Rorke

Department of Information and Computer Science
University of California, Irvine
Irvine, CA 92717

Inference and Learning

Most machine learning methods involve some form of *induction* or inference from specific to general statements, consequently they are often called "data-driven", "empirical", or "similarity-based" learning methods (see, e.g. (Quinlan 1979). Recently, attention has been given to a complementary class of "knowledge-driven", "analytical", or "explanation-based" (EBL) learning methods (see, e.g., DeJong 1988), but these methods have been characterized in terms of *deduction*. There is a third form of inference called *abduction*, and it is argued that abductive inference is at least as fundamental and important as inductive and deductive inference.

I claim that the models of EBL proposed in (Mitchell 1986) and even in (DeJong 1986) and (O'Rorke 1986) should be viewed as *first attempts* at capturing the informal idea of EBL. Intuitively, EBL is "learning based upon explanations." So it seems reasonable to expect EBL theories and systems to include some component aimed at describing or implementing processes for constructing explanations. Weaknesses in the explanation component may reasonably be viewed as weaknesses in EBL formalizations or implementations.

I claim that the initial attempts at formalizing and implementing EBL can be improved upon by introducing more sophisticated models of the explanation process. A good first step in this direction is to view explanation as a kind of plausible inference process — *one that is not often deductive.* The particular forms of plausible explanatory inference explored here are based upon Charles Sanders Peirce's notion of abduction.

In the following sections, I first attempt to be a bit clearer about my usage of the term abduction. Then I briefly describe one early model of EBL and two more recent EBL methods integrating abduction and learning. I argue that incorporating improved abduction methods yields specific improvements over the early model of EBL.

Abductive Inference

Since Peirce's time, a great deal of work has been done on explanations and abduction. This work has taken place both outside of AI in fields such as philosophy (Harman 1965; Peirce 1958; Thagard 1981), and psychology (Donaldson 1986) and within AI in research areas such as expert systems, naive physics, and natural language comprehension (Charniak 1986; Josephson 1987; Morris 1987; Pople 1973; Reggia 1984; Reiter 1987; Schank 1986). Within AI the term abduction is probably being used more broadly than Peirce originally intended. In many AI tasks, it must be decided which of several possible explanations is the best one. In these situations, it is often necessary to evaluate competing explanations. AI researchers often use the term abduction to mean something roughly equivalent to Harman's *inference to the best explanation* (Harman 1965). In other words, in AI the term abduction is often used so as to cover both the generation and evaluation of explanations. I go along with this trend: whenever I speak of abduction informally I mean any method for generating and evaluating explanations.

A survey of the AI literature reveals a number of different proposals for automating abduction. I focus on a particular kind of automated abduction closely related to ideas of Peirce and Hempel. Peirce (1958) used the term abduction as a name for one particular form of explanatory hypothesis generation. His description was basically: "The surprising fact C is observed; But if A were true, C would be a matter of course, hence there is reason to suspect that A is true." Hempel (1965) suggested viewing some explanations as deductive arguments where the thing to be explained follows from a set of general laws and specific facts. Hempel called explanatory accounts of this kind "explanations by deductive subsumption under general laws, or *deductive-nomological (D-N) explanations.* (The root of the term *nomological* is the Greek word *nomos* for law.)" The versions of abduction focused on here combine these ideas as follows. As a first approximation, Peirce's "C is a matter of course if A is true" is represented as "A im-

plies C." Observations C are explained in terms of laws such as "A implies C" and facts or hypotheses such as A. Abduction attempts to reduce observations to known facts by repeatedly backward chaining on laws cast as logical implications.

It is a commonly held misconception that "deductive abduction" is an oxymoron because deduction and abduction are fundamentally incompatible. They are compatible in the sense that deduction may serve abduction: when something is shown to be true, the process of deduction usually supplies a proof that may be considered to be an explanation of why the conclusion is true.

However, deduction fails when a deductive procedure cannot find a proof (explanation) of a conjecture (or observation) from a given set of facts. Abduction generally does not simply fail when no explanation of a given observation can be found from given facts. Instead, abduction often involves making new assumptions. For example, in order to explain observed symptoms, a physician or a medical expert system may assume that a patient has an infection, even though the infection has not been observed, perhaps because it is internal.

Pople's (1973) mechanization of abductive logic goes beyond deduction and provides an EBL model or system with a limited capability for making assumptions in order to complete explanations. Pople's abduction method includes a *synthesis* operator which merges hypotheses, assuming the unified result.

The major problem with this early mechanization of abduction is that it does not address issues that arise when there are many competing explanations. How does one avoid a combinatorial explosion of possibilities while searching for plausible explanations? How does one weigh the evidence and decide that one explanation is more plausible than another?

One class of approaches to these problems involves introducing scoring functions that assign numeric "costs" to potential (partial) explanations. For example, Stickel (Hobbs et al. 1988) has suggested a heuristic approach to evaluating explanations in the context of natural language processing. O'Rorke and his students have experimented with a best first heuristic search program named AbE using several different heuristic scoring functions in the context of physical (O'Rorke, Morris, & Schulenburg 1989) and psychological (Cain, O'Rorke & Ortony 1989) explanations.

Combining Abduction and Learning

In the influential model of EBL presented by Mitchell et al (1986) and in the implementation presented in (Kedar-Cabelli 1987), learning is based upon explanations generated by a deductive theorem prover. The learning method is essentially a form of lemma caching or deductive macro-learning.

This form of EBL has been criticized on the grounds that it only improves efficiency and does not involve "learning at the knowledge-level" (as defined by Dietterich (1986)). The deductive closure of the knowledge-base does not change as a result of learning because the macro-learning method specializes existing general knowledge, even though it generalizes given examples.

This early model of EBL rests on a purely deductive model of abduction. Integrating more sophisticated models of the explanation process with learning leads to interesting new models of EBL.

For example, a new model of EBL can be had by integrating Pople's mechanization of abduction with the usual EBL macro learning procedure. An implementation based on this idea called AMAL was reported in (O'Rorke 1988). AMAL used abductive inference to "leap to conclusions" during the process of explaining an observation. While the macro-learning component of AMAL did not contribute new knowledge, AMAL's epistemic state could change because assumptions were often needed in order to explain observations. Adding these assumptions changed the knowledge base. This improved upon the EBG version of EBL by allowing a limited form of learning at the knowledge level. However, the assumptions made by AMAL were typically very specific statements closely related to the observation being explained in given examples.

AMAL also suffered from the weakness of Pople's initial mechanization of abduction, namely its inability to evaluate alternative explanations. Including methods for evaluating explanations leads to more powerful combinations of abduction and learning. The abduction engine AbE, a PROLOG meta-intepreter originally based on AMAL, does heuristic search for plausible explanations. AbE is now in use in several case studies of abduction and learning. One such study is aimed at exploring the possibility that abduction can provide a handle on how one might automate massive changes in systems of beliefs. O'Rorke, Morris & Schulenburg (1989) sketches this case study based on an episode in the history of science known as the chemical revolution. The learning process of interest in this study is not macro-learning. It is a theory revision or knowledge-base refinement process. The process starts out with an incorrect theory or knowledge-base and is confronted with an anomaly, an observation that contradicts a prediction of the initial theory. Abduction is used to explain the anomalous observation and to form hypotheses corresponding to crucial parts of a revised theory. In this study, abduction contributes to knowledge level learning of very general theoretical statements.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Award No. IRI-8813048 and supported in part by an Irvine Faculty Fellowship from the University of California, Irvine Academic Senate Committee on Research. This work has been influenced by interactions with Steven Morris, Pat Langley, Donald Rose, Tim Cain, David Aha, Stephanie

Sage, David Schulenburg, and other members of the AI and machine learning community at the University of California, Irvine.

References

Cain, T., O'Rorke, P. & Ortony, A. (1989). Learning to Recognize Plans Involving Affect. *Proceedings of the Sixth International Machine Learning Workshop.* (pp. 209-211) Ithaca, NY: Morgan Kaufmann, Publishers, Inc.

Charniak, E., McDermott, D. (1986). *Introduction to Artificial Intelligence.* Reading, MA: Addison-Wesley Publishing Company.

DeJong, J. & Mooney, R. (1986). Explanation-Based Learning: An Alternative View. *Machine Learning* 1:2, 145-176.

DeJong, J. (Ed.) (1988). *Proceedings of the American Association for Artificial Intelligence Spring Symposium on Explanation-Based Learning* Stanford, CA: AAAI.

Dietterich, T. (1986). Learning at the Knowledge Level. *Machine Learning* 1:3, 287-316.

Donaldson, M. (1986). *Children's Explanations: A Psycholinguistic Study.* Cambridge: Cambridge University Press.

Harman, G. (1965). The Inference to the Best Explanation. *Philosophical Review* 74, 88-95.

Hempel, C. (1965). *Aspects of Scientific Explanation, and Other Essays in the Philosophy of Science* New York, New York: MacMillan.

Hobbs, J., Stickel, M., Martin, P., & Edwards, D. (1988). Interpretation as Abduction. *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics* (pp. 95-103) Buffalo, NY: ACL.

Josephson, J., Chandrasekaran, B., Smith Jr., J., & Tanner, M. (1987). A Mechanism for Forming Composite Explanatory Hypotheses, *Institute of Electrical and Electronics Engineers Transactions on Systems, Man and Cybernetics, Special Issue on Causal and Strategic Aspects of Diagnostic Reasoning* 17:3, 445-454.

Kedar-Cabelli, S., & McCarty, L. (1987). Explanation-Based Generalization as Resolution Theorem Proving, *Proceedings of the Fourth International Workshop on Machine Learning,* (pp. 383-389). Irvine, CA: Morgan Kaufmann.

Mitchell, T., Keller, R., & Kedar-Cabelli, S. (1986). Explanation-Based Generalization — A Unifying View. *Machine Learning* 1:1, 47-80. Hingham. MA: Kluwer.

Morris, G., Finin, T. (1987). *Abductive Reasoning in Multiple Fault Diagnosis,* (MS-CIS-87-63, LINC Laboratory 78). Philadelphia, PA: Department of Computer Science, University of Pennsylvania.

O'Rorke, P. (1986) *Explanation-Based Learning via Constraint Posting and Propagation,* (Ph.D. Thesis) Urbana, IL: Department of Computer Science, University of Illinois.

O'Rorke, P. (1988). *Automated Abduction and Machine Learning.* Irvine, CA: Information and Computer Science Department, University of California, Irvine.

O'Rorke, P., Morris, S., & Schulenburg, D. (1989). Abduction and World Model Revision. *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society* (pp. 789-796). Ann Arbor, MI: Lawrence Erlbaum and Associates.

Peirce, C. S. S. (1931-1958). *Collected Papers of Charles Sanders Peirce (1899-1914).* C. Hartchorne, P. Weiss, & A. Burks (Eds.), Cambridge, MA: Harvard University Press.

Pople, H. E. (1973). On the mechanization of abductive logic. In *Proceedings of the Third International Joint Conference on Artificial Intelligence* (pp. 147–152) Stanford, CA: Morgan Kaufmann.

Quinlan, R. (1979). Discovering Rules by Induction from Large Collections of Examples. In Donald Michie (Ed.) *Expert Systems in the Micro Electronic Age.* Edinburgh: Edinburgh University Press

Reggia, J. A., & Nau, D. S. (1984). An abductive non-monotonic logic. In *Proceedings of the Workshop on Non-Monotonic Reasoning* (pp. 385–395). New Palz, N.Y.: American Association for Artificial Intelligence.

Reiter, R., de Kleer, J. (1987). Foundations of Assumption-Based Truth Maintenance Systems: Preliminary Report, (pp. 183-188). *Proceedings of the National Conference on Artificial Intelligence.* Seattle, WA: Morgan Kaufmann, Publishers, Inc.

Schank, R. (1986). *Explanation Patterns: Understanding Mechanically and Creatively.* Hillsdale, NJ: Lawrence Erlbaum and Associates.

Thagard, P. (1981). Peirce on Hypothesis and Abduction, (pp. 271-274). *Proceedings of the C. S. Peirce Bicentennial International Congress.* Lubbock, Texas: Texas Tech University Press.

An Approach to Theory Revision Using Abduction

Steven Morris Paul O'Rorke

Department of Information and Computer Science
University of California, Irvine
Irvine, CA 92717

Revising theories using abductive hypothesis formation

This extended abstract sketches an approach to theory revision using abductive hypothesis formation. The need for theory revision is typically recognized when a theory is found to be in contradiction with new observations. The task is then to determine what revisions will result in a new theory that is in accord with observation. Most approaches to theory revision involve direct transformations producing the new theory from the original "old" theory. These transformations are generally very much like "editing" or "patching". Two combinatorial problems occur in these transformations. The first involves the identification of the erroneous subset of the original theory. The second involves the identification of the correct changes in the erroneous parts of the original theory. In some situations, these combinatorics are likely to overwhelm editing approaches to theory revision.

There seems to be some psychological evidence that people sometimes do not do this sort of editing. In Shrager and Klahr's "instructionless learning" experiments, subjects were asked to "figure out" devices such as the BigTrak toy programmable tank. Shrager (1987) comments:

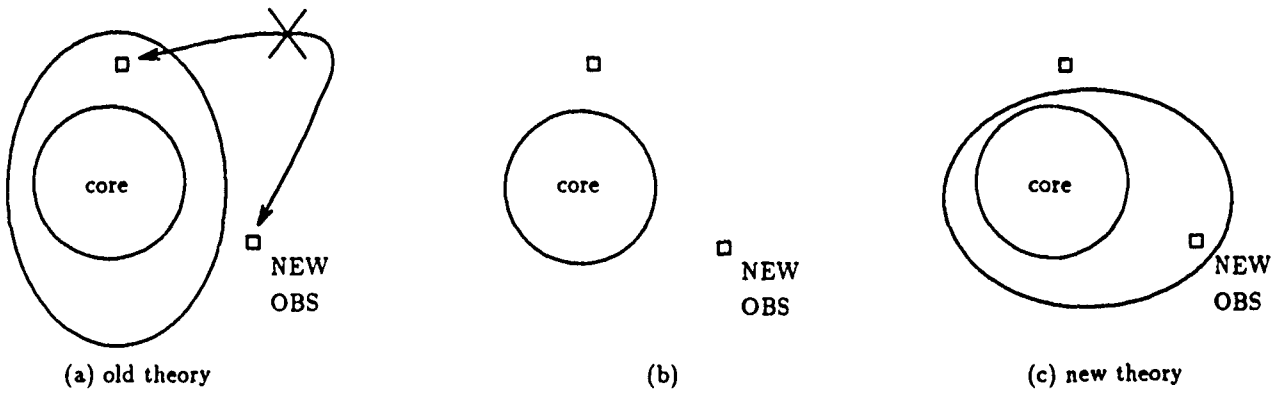
> *... we observed that between interactions with the BigTrak, subjects changed their theory of the device. A number of empirical generalizations seem to hold about the nature of these changes... Instead of trying to determine in detail what led to a failed prediction, subjects usually observed what (positive behavior) took place and changed their theory according to that observation...*

When a surprising observation contradicts a prediction of the original theory, the approach to theory revision explored in the present paper involves retracting questionable beliefs. However, it is not necessary to start by trying to identify an individual incorrect belief or even a small set of culprits. Instead, the approach explored in this paper assumes that the initial theory has some internal structure and that more general fundamental principles can be separated from relatively specific, less basic statements. A "core" subset of the original theory, a set of basic statements having nothing to do with the anomaly, is retained while less central beliefs are suspended. Then the unexpected new observation is explained in terms of the remaining, relatively solid basic principles. As we will see in the example presented, this explanation process can generate hypotheses, suggesting extensions to the basic theory that will result in proper explanation of the new observations.

This approach to theory revision is sketched in Figure 1 using Venn diagrams. In the first stage (a) of theory revision an anomaly is noted. A new observation contradicts a prediction of the old theory, as indicated by the X linking a point in the old theory and a point outside of it. In the next stage (b) the old theory is reduced to the core subset.[1] Starting from this subset, an explanation of the new observation is abduced with hypotheses being introduced in the process. These hypotheses then form the basis for extensions to the core theory resulting in a new theory (c). This revised theory no longer makes the erroneous prediction of the old theory.

We do not explore here the initial step of falling back on basic principles and shrinking the original theory. Instead, we focus on the step from Figure 1(b) to Figure 1(c). We concentrate on the claim that the process of explaining unexpected new phenomena can lead by

[1] Notice that neither the prediction nor the surprising observation are included in the reduced core subset of the original theory. The circles and ellipses designate theories closed under deductive inference. The figure captures the notion that neither the prediction nor the contradictory observation should be implications of the core theory.



(a) old theory (b) (c) new theory

Figure 1: Theory revision using abduction for hypothesis formation.

abductive inference to new hypotheses which can form crucial parts of new theories. We then discuss some of the issues involve in arriving at a new, revised theory.

A case study in chemistry

Let us consider in more detail some aspects of the above theory revision framework by looking at a case study of the chemical revolution. Using the language of Qualitative Process theory (Forbus, 1984) we have encoded into rules and facts a domain theory, PT, that captures a portion of the phlogiston theory of the Middle Ages, including some basic knowledge concerning complex substances, and some key laws of QP theory. Using an abductive inference system named AbE, we have done a case study of the shift from the phlogiston theory to the oxygen theory (O'Rorke, Morris, & Schulenburg, in-press).

Phlogiston theory was developed to explain, among other things, the phenomenon of combustion. It explains combustion as an outflow of a component called phlogiston from the combusting material. The theory predicts that a combusting piece of substance loses weight due to this outflow.[2]

Figure 2 shows a generalized explanation of weight loss during combustion using our encoding of PT. This explanation is represented by an AND-tree with each line of the figure showing one tree node. The children of a node are indicated by equally indented lines following the node. For example, the nodes 'member(amt-in...)' and 'amt(S) =...' are siblings. Each leaf of the tree is followed by a box indicating that it is an hypothesis H, a background fact F, or a case fact CF. (Cases facts are facts, usually observations, that are relevant to the observation being explained.)

The root of the tree is the observation that the weight of a combusting piece of some substance, S, is decreasing. This is expressed as: The time derivative's sign, ds, of weight(S) is neg(ative). The nodes beneath the observation have been generated by backchaining on rules, and unification with facts. *Qprop* stands for 'qualitatively proportional'. Qprop(x,y,pos) means that as x increases, y increases, and vice versa; and qprop(x,y,neg) means that as x increases, y decreases, and vice versa. *Phlog* stands for phlogiston. _72 is a Skolem constant that represents an unspecified list of amounts of components of S. These would be the portions of S left after burning.

The explanation in Figure 2 is intended to reflect the kind of generalization a scientific theory would assert. It is based on specific explanations of specific combusting substances that would have been observed and explained by PT. Such generalization is one means by which a theory predicts. This generalized explanation predicts that *any* combusting substance loses weight. For example, when told that a quantity of phosphorus is combusting, PT predicts that the phosphorus is losing weight. The explanation would be that of Figure 2 instantiated with S = phosphorus.

PT *predicts* weight loss in that, if all the leaves of the explanation in Figure 2 are grounded in fact, the weight loss follows deductively. This generalized explanation is a schematic proof of weight loss, and perhaps the leaves should be referred to as *potential* facts. It should be noted that the various facts at the leaves of this explanation have different statuses. For a particular instantiation of S, the facts process(combustion) and active(combustion,S), asserting that S is undergoing combustion, would be grounded in observation, whereas a "fact" such as component(phlog,S) would not be an

[2]Although the phlogiston theorists may not have originally taken weight into account, we extend our encoding PT to include weight considerations.

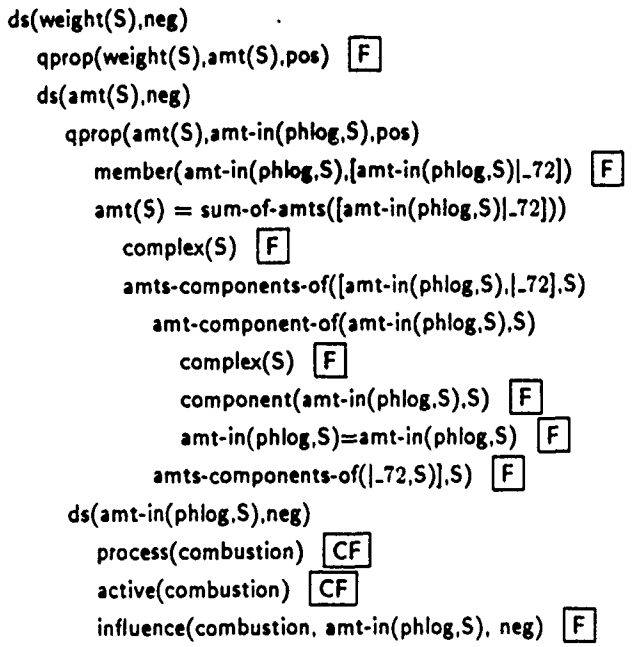


Figure 2: Why weight of burning substance S decreases.

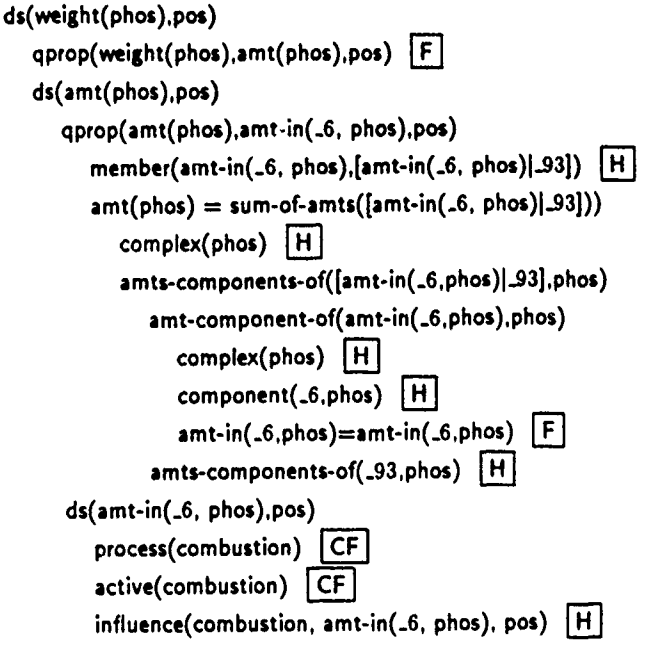


Figure 3: Why weight of burning phosphorus increases.

observation, but rather a theoretical assertion of PT. However, we will continue to refer to literals like component(phlog,S) as facts (*á* la PROLOG).

Generating new theory elements

We now consider the example of burning a piece of wood. The weight of the wood before burning will be greater than the weight of the ash left after burning. The phlogiston theorists's model of this combustion would be that wood = phlogiston + ash, and that the combustion is the outflow of phlogiston. For a phlogiston theorist, the decrease in weight would make sense.

Today we know that the burning of wood is a much more complex process. Not only are some of the formed oxides missing in the ash (for example, carbon monoxide), but also, due to the heat of combustion, other weighty components such as water escape in gaseous form. Thus the weight increase due to oxidation is confounded by other losses which result in an apparent weight decrease when looking only at the residual ash. However, if one burns an elemental substance such as phosphorus, such a confoundment does not occur. A combusting piece of phosphorus gains weight.

Our system AbE, using PT as a domain theory, is presented with the observation that a combusting quantity of phosphorus is *gaining* weight, Assume that some mechanism has detected the contradiction between this observation and PT's prediction of weight loss. Further assume, that as a result, AbE falls back to a core subset of PT comprising the basic laws about Qualitative Processes and the basic laws about complex substances (a complex substance has components; the amount of a complex substance is the sum of the amounts of its components; etc.) This core theory excludes the law that states that combustion causes a decrease in the amount of phlogiston in a combusting substance:

influence(combustion,amt-in(phlog,S),neg).

AbE is now asked to explain the new contradictory observation using only this subset. AbE does this by attempting to generate an explanation that reduces the observation to the given facts. Failing this, AbE generates an explanation that has some hypotheses at its leaves.

The explanation produced by AbE is shown in Figure 3. It states that some some hitherto unknown component _6 of the piece of combusting phosphorus has increased in amount, and is thus responsible for the overall increase in weight. In summary, a new explanation of combustion involving an increase of a component rather than a decrease of a component is proposed. AbE abductively generates the hypotheses that the piece of combusting phosphorus is a complex substance, and that it contains a component, _6, the amount of which increases during combustion. This hypothesized new component may be interpreted as corresponding to oxygen. This demonstrates that abduction can be used to form hy-

potheses corresponding to essential parts of new theories. A generalized version of the explanation in Figure 3 could be proposed producing new theory components such as influence(combustion, amt-of-in(_6,S), pos).[3] These generalized components, along with the core theory, would provide a theory of combustion that predicts weight increase for any combusting substance.

Revising the theory

There are many possible reactions within a scientific community to a new contradictory observation. These range from questioning the observation to taking the new observation as a sign that a current theory is flawed and in need of revision. We discuss here this latter course.

In the above combustion example, new theory components are hypothesized, which, in conjunction with the core theory, explain the new observation. However, the process of falling back to the core theory may have thrown out parts of PT that are not responsible for the prediction of weight loss in combustion. In order to determine which components of PT should be blamed, one may compare the generalized explanation of Figure 2 and the specific explanation of Figure 3 to determine differences between each explanation that arise from the non-core theory components in each theory. Doing so identifies two such discrepancies between the explanations:

- component(amt-in(phlog,S),S) 1(a)
  vs. component(amt-in(_6,phos),phos) 1(b)

- influence(combustion,amt-in(phlog,S),neg) 2(a)
  vs. influence(combustion,amt-in(_6,S),pos). 2(b)

Neither pairs of assertions are necessarily contradictory. However, the close parallel between the two explanations suggests that these pairs of theory elements have similar roles in their respective models of combustion. On this basis, one may consider 1(a), 2(a), and their specializations as candidates for excision. For example: component(amt-in(phlog,wood), wood) and influence(combustion,amt-in(phlog,wood),neg). Other non- core components of PT not blamed in this comparison should be considered for inclusion in the new theory. Thus a new theory, NT, may be obtained as: core + hypothesized components + unblamed old components.

The above revision procedure produces a candidate new theory that may be capable of explaining the new observation of weight gain during combustion. If one assumes that the new observation has been checked and is considered accurate, one is still left with the question of the prior observations of weight *loss* during combustion. Clearly NT as it stands can not explain these observations.

One of the theories can prevail over the other if it can demonstrate that the other theory is misapprehending the phenomena it purports to explain. The wood burning case with its missed confounding influences and missed phenomena provides an example of such misapprehension. Arriving at an expanded NT that explains burning wood requires reasoning along the lines of: "Let us believe the new model of combustion. Perhaps there are confounding influences present in the old combustion events. The simplest model explaining these observations would involve _6 entering and some other substance leaving (perhaps even phlogiston?) with a net decrease in weight. But what process could be responsible for the departing substance?"

Arriving at a suitable hypothesis for such a confounding process will usually require experiment and the gathering of more observations that allow the proposal of other active processes. By manually making such new data available to our system, we hope to model the subsequent hypothesis generation that would result in explanation of prior observations that are in conflict with the new, revision-provoking observation.

When one admits confounding influences, the door opens for arbitrarily complex theorizing. One would like to entertain hypotheses in a conservative manner. One would like a heuristic that strives for simple processes and a minimal number of them. For example, simplicity argues for trying a model of combustion that involves the flow of only one substance, and not a model that involves the flow of two substances in opposite directions, with one flow dominating the other with respect to weight change.

Thus one might consider either of two models, in which weight change discrepancies are explained by a second process acting in one subset of the observations:

(1) Combustion = outflow of phlogiston. Weight loss observations are due to combustion only. Weight gain observations are due to combustion plus a heavier inflow process.

(2) Combustion = inflow of _6. Weight gain observations are due to combustion only. Weight loss observations are due to combustion plus a heavier outflow process.

Heuristics that propose such minimal revisions are necessary to reduce the combinatorics of hypothesis generation. Another potential and difficult problem is that the new theory components added to NT might combine with old components remaining in NT, such that contradictions may be deduced.

[3] AbE does not currently perform this generalization.

Editing versus deep revision

One may view various types of theory revision as lying on a spectrum ranging from minor corrections and adaptations of theory to deep, revolutionary change, as exemplified by major scientific shifts. We suggest that the above method of theory revision is appropriate for situations in which substantial changes are required. An editing approach to theory revision is at the other end of the spectrum. Such an approach is more oriented toward a theory which is slightly incorrect or incomplete, but can be slightly modified to explain new observations.

One difficulty that editing approaches may have is introducing new relations between objects in a principled way. Edits that revise relationships between objects, or that introduce new objects, have a relatively small chance of being correct. Consequently, a large number of candidate revisions may need to be introduced and tested. Even then, there is no guarantee of proposing the right edit without guidance from first principles. On the other hand, theory driven approaches to revision, such as ours, can use relatively solid knowledge to guide the revision process, and thus stand a better chance of hypothesizing appropriate new relations and objects.

We view these two approaches to revision as complementary. If one has a theory that is essentially adequate, then the editing approach may be a useful technique for arriving at a more finely tuned, final theory. However, when the theory is very wrong, falling back to a core theory may be a necessary prelude to the type of theory revision processes needed to create new theoretical entities. We consider detecting which method of theory revision is appropriate to be an interesting problem.

Conclusion

Theory revision can profitably be viewed as a process that involves hypothesis formation by abduction. When a new observation contradicts a prediction of the theory, one approach is to suppress questionable details of the original theory and to derive an explanation of the observation based on more solid, basic principles of a core subset of the theory. The abductive generation of this explanation can lead to new hypotheses that can form crucial parts of a new theory. Comparison of the explanation of the new observation, to an explanation of the contradictory prediction under the old theory, can provide a focus for blame assignment. A candidate new theory results from a conjunction of the core theory, the new hypotheses, and the unblamed non-core theory. However, re-explaining old observations may require more sophisticated revision involving interacting processes.

Related work

Falkenhainer (1988) and Rajamoney (1988) describe approaches to theory revision in QP domains, including constrained hypothesis generation and the role of experimentation. Thagard (in-press) examines the conceptual changes that occurred during the overthrow of the phlogiston theory, and gives a conceptual map of several stages of the transition.

Acknowledgments

Ideas in this paper have evolved in discussions with members of the machine learning community at the University of California, Irvine. Special thanks to Pat Langley, Deepak Kulkarni, and Don Rose for discussions of scientific discovery. Discussions with Paul Thagard on Peirce, abduction, the chemical revolution, and scientific revolutions were inspirational. This paper is based on work supported in part by an Irvine Faculty Fellowship from the University of California, Irvine Academic Senate Committee on Research and by grant number IRI-8813048 from the National Science Foundation.

References

Falkenhainer, B. (1988). *Learning from physical analogies: A study in analogy and the explanation process.* Ph.D. thesis (Report Number UIUCDCS-R-88-1479). Urbana-Champaign, IL: University of Illinois, Department of Computer Science.

Forbus, K. D. (1984). Qualitative process theory. *Artificial Intelligence, 24*, 85–168.

O'Rorke, P., Morris, S., & Schulenburg, D. (in-press). Theory formation by abduction: A case study based on the chemical revolution. *Proceedings of the Symposium of Computational Models of Scientific Discovery and Theory Formation.*

Rajamoney, S. (1988). *Explanation-based theory revision: An approach to the problems of incomplete and incorrect theories.* Ph.D. thesis (Report Number UILU-ENG-88-2264). Urbana-Champaign: University of Illinois, Coordinated Science Laboratory.

Shrager, J. (1987). Theory change via view application in instructionless learning. *Machine Learning, 2*, 247–276.

Thagard, P. (in-press). The conceptual structure of the chemical revolution. *Philosophy of Science.*

Learning from Examples and an "Abductive Theory"

William W. Cohen

Rutgers University
Department of Computer Science
New Brunswick, NJ 08903

Summary. Deductive and abductive reasoning use a common representation of knowledge: a theory of the world that allows reasoning from causes to effects. This paper describes a system which performs *inductive* reasoning using knowledge represented in a similar form. The benefit of this approach is that a common knowledge base can be used for inductive, abductive and deductive reasoning. The similarities and differences between this learning system and abductive reasoning systems are discussed.

Introduction

There are three types of reasoning: *deduction*, *abduction*, and *induction*. Broadly speaking, these types of reasoning can be described as reasoning from causes to effects, from effects to probable causes, and from specific facts to general cause-effect relationships, respectively.

The knowledge used for deductive reasoning is almost always represented in "cause-to-effect" form: that is, a theory which describes the effects of various underlying causes. While much of the earlier work on abductive reasoning used knowledge in "effect-to-cause" form, most recent work has concentrated on abductive reasoning techniques that use knowledge in "cause-to-effect" form as well (this problem has been called "diagnosis from first principles" or "model-based diagnosis".) It is generally agreed that knowledge in "cause-to-effect" form is easier to acquire and maintain.

Inductive reasoning systems, in contrast, rarely represent knowledge about the world in any explicit form. The prototypical inductive reasoning task is *concept learning*, the problem of finding an unknown concept given positive and negative examples of members of that concept. In order to make concept learning possible at all, generally a concept learning system must make some assumptions about the form of the unknown concept. These assumptions are usually syntactic constraints on the form of the concept to be learned; for instance, that it be expressed as a conjunction of certain features, or as a well-balanced decision tree. These constraints are typically the only sort of world knowledge available to a concept learner.

This paper describes a concept learning system which learns using examples and knowledge about the world written in cause-to-effect form. Due to the similarities between this form of knowledge and the knowledge used by model-based diagnosis systems, I will call a theory written in the appropriate format an "abductive" theory. One benefit of this approach is that a common knowledge base can be used simultaneously (at least in principle) for inductive, abductive and deductive reasoning. Another benefit is that knowledge in this format is easier to acquire and maintain.

Via a short case study, I will show that this can also be an appropriate and natural way of representing knowledge for an inductive learner.

Space constraints preclude inclusion of proofs or detailed experimental results. The interested reader is referred to (Cohen 1989).

Statement of the problem

Preliminary Definitions

Consider a Horn clause theory Θ. Let $\mathcal{O}$ be the set of all possible ground atomic formula in the Herbrand universe of Θ. I will call these formula *possible observations*. The *target concept* is some set $T \subseteq \mathcal{O}$. The target concept represents the set of observations which are "true" in the real world. An *example of* T is an element of $x \in \mathcal{O}$, labeled with "+" if $x \in T$ and "–" if $x \notin T$.

Let p_x denote any AND-OR proof of the formula x

in the theory Θ, and let a_x be the generalized version of that proof (the "explanation structure" for the proof) obtained by using Mitchell's goal-regression algorithm (Mitchell, Keller, & Kedar-Cabelli 1986). I will call a_x *valid* if $\forall y$, (y provable using a_x) $\Rightarrow y \in T$. Finally, a theory is called *abductive for* T if $\forall x \in T$, $\exists a_x$: a_x is valid.

Intuitively, each proof in an abductive theory is a "tentative explanation" of the observation which it proves, and the generalized proof represents the "chain of reasoning" used to produce the proof. These chains of reasoning can be either "valid" — *i.e.*, that same chain of reasoning always holds — or "invalid" — *i.e.*, that same chain of reasoning can sometimes be used to support a conclusion which conflicts with reality. The crucial property of an abductive theory is that for every true observation, there is some valid explanation — in other words, one of the explanations suggested by the theory is correct.

This definition seems to fit many cases of practical interest. For example, consider a theory that involves some element of plan recognition. The goals of one or more agents are typically unknown and must be assumed. Often, any of several assumptions which could be introduced would suffice to explain an action, but only explanations based on the correct assumption will be valid.

The definition also fits theories for which are not normally considered abductive. For example, consider an abductive theory Θ for opening bids in the game of contract bridge. One of the predicates defined in this theory might be the predicate *opening_bid*($Hand, Bid$), where *Hand* is a term describing a bridge hand, and *Bid* describes a possible opening bid.

If this theory is abductive, it need not be correct, in the sense that some of the *opening_bid* goals provable in the theory might be for incorrect opening bids. For instance, the theory might contain some overly general, heuristic rules for how to make opening bids, but might not contain knowledge about when these heuristic rules should be applied. So given a bid, the approximate theory could not be used to to determine if the bid was correct or not, but could be used to construct tentative explanations of why the bid was made. If the theory is abductive for the target concept of "correct opening bid", then for each correct bid, one of these explanations is always valid.

The goal of learning

Inductive learning and abduction are very similar tasks. In both cases, the problem is to come up with a hypothesis which explains a particular set of observed effects. Ideally, the hypothesis should exactly coincide with the true underlying causes of the effects; however, this ideal goal is not attainable except in trivial circumstances. Learning systems and abductive reasoning systems differ in how this goal is relaxed. Abductive reasoning systems typically produce as output the set of *all possible* hypotheses which satisfy some relatively weak definition of minimality (for instance, minimality under the partial order of set inclusion). The criterion of success is whether this set contains all of the most likely hypotheses. Learning systems, in contrast, usually produce a *single* hypothesis explaining a set of phenomena, using a relatively strong definition of minimality (usually small syntactic size, relative to a particular encoding). Various measures of success have been proposed for learning systems. The learning goal used in this paper is Valiant's criterion of probably approximately correct learning (Valiant 1984), in which a learner succeeds if *future predictions* made by the hypothesis are correct in a probabilistic sense.

More precisely, let $\mathcal{T}$ be a space of possible target concepts, let $T \in \mathcal{T}$ be a target concept, let D be a probability density function, and let *size*(T) be some complexity measure on concepts. Define the error of a hypothesis H with respect to T and D to be $error(H, T, D) \equiv D(T - H) + D(H - T)$. A learning algorithm LEARN is a function which takes as input a sample of T containing m examples and outputs a hypothesis: that is, a guess as to what T is. A learning algorithm is said to be *polynomially probably approximately correct* for $\mathcal{T}$ if there is some polynomial function $m(1/\epsilon, 1/\delta, n)$ such that for any probability distribution function D, for any $T \in \mathcal{T}$

1. LEARN runs in time polynomial in its inputs
2. For a sample S of size $m(1/\epsilon, 1/\delta, n)$ of some target concept T such that *size*(T) $\leq n$,

$$Prob(error(LEARN(S, T, D)) > \epsilon) < \delta$$

In other words, LEARN *probably* — with probability at least $(1 - \delta)$ — returns an *approximately correct* hypothesis — a hypothesis with error less than ϵ — and is constrained to run in time *polynomial* in $1/\epsilon$, $1/\delta$, and the size of the target concept.

The function $m(1/\epsilon, 1/\delta, n)$ is called the *sample complexity* of the function LEARN; it indicates how many examples are needed to ensure that the hypothesis is probably approximately correct.

Note that the error is defined with respect to the same probability density function D from which examples were drawn. This can be interpreted as saying that the accuracy of the hypothesis produced by the learner is guaranteed only for the same population from which the training examples were drawn.

A learning algorithm

The algorithm

A simple learning algorithm is the following.

Algorithm A-EBL(S):

1. Enumerate all the generalized proofs $a_{x1}, \ldots, a_{xr}$ of the positive examples in the sample S.
2. Discard those a_{xi}'s which can be used to prove some negative example.
3. Use a greedy algorithm to find a minimal subset COV of the remaining a_{xi}'s such that every positive example $x+$ can be proved by some $a_{xi} \in COV$. The greedy algorithm always adds to COV an a_{xi} that maximizes the ratio of the number of uncovered examples to the size of a_{xi}.
4. Return the hypothesis

$$H = \{x : x \text{ is provable with some } a_{xi} \in COV\}$$

The *size* of a generalized proof a_x is defined to be the number of nodes in the proof tree. Further details of the algorithm can be found in (Cohen 1989).

Formal Analysis

It can be shown that this algorithm satisfies the learning goal described in the previous section, and that its sample complexity nearly optimal.

Let $\mathcal{T}_\Theta$ denote the set of all specializations T of a domain theory Θ such that $T = G_1 \cup \ldots \cup G_k$, where $\forall i$, $1 \leq i \leq k$, there is some proof p_{x_i} such that G_i is the output of $EBG(p_{x_i})$. In other words, $\mathcal{T}_\Theta$ is the set of all possible target concepts T which are explainable by a some set of generalized proofs: in the terminology introduced above, T is the set of all target concepts such that Θ is abductive for T. Let the size of $T \in \mathcal{T}_\Theta$ be the sum of the sizes of the abstract explanations which define T; finally, let $|\Theta|$ denote the number of clauses in Θ.

In (Cohen 1989) is a proof of the following theorem.

Theorem 1 *A-EBL(S) is a polynomial probably approximately correct learning algorithm for $\mathcal{T}_\Theta$ with sample complexity of*

$$m(\frac{1}{\epsilon}, \frac{1}{\delta}, n) = O(\max(\frac{1}{\epsilon}\log\frac{1}{\delta}, \frac{n\log|\Theta|}{\epsilon}(\log\frac{n\log|\Theta|}{\epsilon})^2))$$

Furthermore, there exist theories Θ such that every *probably approximately correct learning algorithm for $\mathcal{T}_\Theta$ must have a sample complexity of at least*

$$m(\frac{1}{\epsilon}, \frac{1}{\delta}, n) = \Omega(\frac{1}{\epsilon}\ln\frac{1}{\delta} + \frac{n}{\epsilon})$$

Discussion of the Algorithm

Abduction is often thought of as finding the best explanation of a set of phenomena. When there are multiple explanations, as is assumed to be the case here, simpler explanations are preferred.

This is precisely the intent of the algorithm given above. It uses the heuristic set cover technique to minimize the complexity and number of explanations considered. The hypothesis output by this algorithm is the set of observations predicted by a disjunctive explanation of the example observations. The formal analysis shows that this technique works whenever the target concept T corresponds exactly to the set of observations predicted by some set of generalized explanations; that is, whenever Θ is abductive for T. In short, A-EBL is very similar to many abductive reasoning systems.

The major difference between A-EBL and abductive reasoning systems is that the goal of A-EBL is different. A-EBL is attempting to produce a hypothesis which will make reasonably accurate predictions on later problems, if these problems are drawn from the same population from which the training samples were drawn. Theorem 1 shows that the simple heuristics used in A-EBL are sufficient to satisfy this goal.

Experimental Results

Theorem 1 shows that effective learning algorithms can be designed which make use of knowledge represented as an abductive theory. It remains to be shown that the knowledge to needed solve real-world learning problems can be expressed as an abductive theory.

As an experiment, an introductory text on bridge play (Sheinwold 1964) was used as a source of background knowledge (in the form of a theory), examples, and test data. Almost all of the rules in the theory were clearly and explicitly presented in (Sheinwold 1964), and could be easily transcribed into a Horn

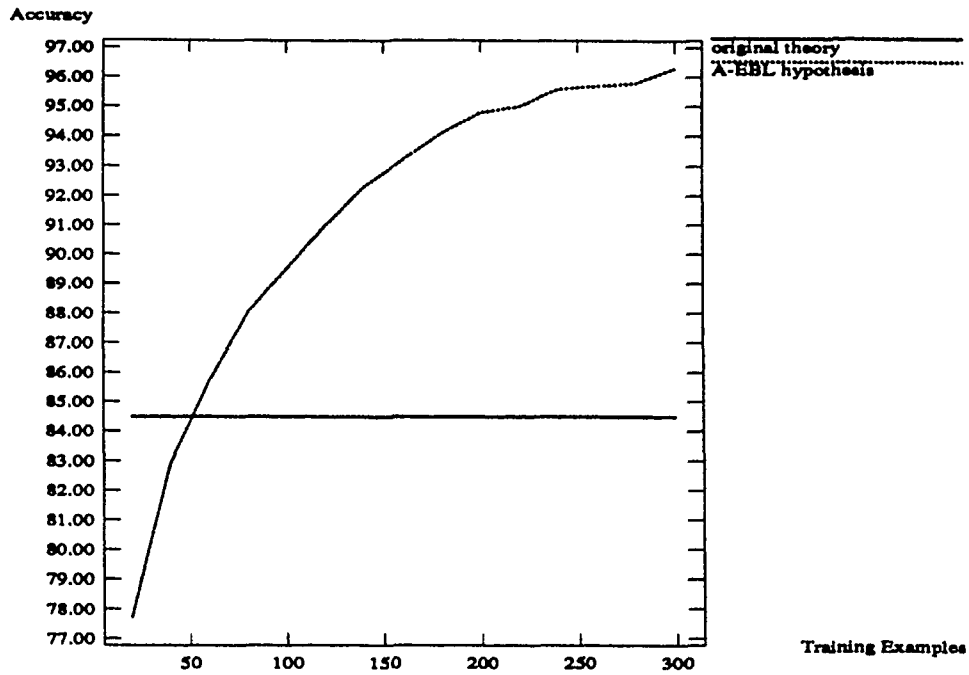


Figure 1: Accuracy of A-EBL's Hypothesis as a Function of Training Set Size

clause theory. However, the resulting theory was not a complete and correct theory of bridge bidding. It is clear (from the accompanying text, if from no other source) that the bidding rules are overly-general. Often the text explicitly states that a rule is merely a heuristic, and should not always be followed. In most of these situations, a series of examples are used to clarify the use of a heuristic rule. The approximate bidding theory can be interpreted, as in the example above, as an abductive theory: each proof can be thought of as a tentative explanation of why a bid might have been made.

The A-EBL algorithm, and a variant of it, was then used to construct a hypothesis for the unknown target concept "good opening bid". This hypothesis was tested using a sample test in (Sheinwold 1964). Both A-EBL and its variant scored well, at 87% correct or above. The original theory Θ scores at only 75% correct; in 25% of the test cases incorrect bids, as well as the correct bids, were proposed by Θ.

An additional experiment was conducted to test A-EBL's behavior on randomly generated training examples. A test set of 1000 hands was randomly generated and classified by the a hand-coded bridge bidding theory. Then a separate training set of 300 hands was generated and classified by the hand-coded theory. A-EBL was then given progressively larger subsets of the training set, and the accuracy of each theory specialization produced was measured by using it to classify the hands in the test set, and comparing the classifications to the correct ones. This experiment was repeated 10 times and the error rates were averaged, using the same test set in each trial. The result is the "learning curve" shown in Figure 1. Performance of the hand-coded theory is also shown for the purpose of comparison. This experiment shows that, as predicted by the theory, A-EBL has good convergence properties on randomly selected data.

In this case, a good argument can be made that the knowledge used for learning was *appropriate*, in the sense that it allowed learning to proceed effectively, and *natural*, in the sense that transcribing this information into usable form was straightforward. Of course, to argue that this is usually (or even often) the case requires many more data points, in the form of other learning problems which can be treated in a similar manner.

Related Work

This work was motivated primarily by the *multiple explanation problem* in explanation based learning, which occurs when explanation based learning techniques are applied to approximate domain theories. Often these approximate theories are abductive theories (at least in some informal sense.) Several other researchers have also addressed this problem. Hirsh (Hirsh 1989) has used the incremental version-space merging (IVSM) method to choose between multiple explanations. However, the IVSM method requires an additional source of information in the form of a *concept description language* which provides an additional bias to the learning system. In (Pazzani 1988), Pazzani discusses mechanisms to choose between alternative explanations in an imperfect domain theory for plan recognition, which is an abductive task. Pazzani identifies five heuristics for selecting explanations; one of these, the heuristic of preferring explanations which account for a larger number of observed changes, bears some similarity to the basic method of A-EBL. Similar heuristics for choosing between multiple explanations in the context of completing an incomplete theory are proposed by Fawcett in (Fawcett 1989). Our work extends these techniques by giving a precise way of weighting the complexity of an explanation and the number of observations that it covers, and justifying this heuristic with a pac-learning analysis.

Conclusion

This paper describes a learning system that uses knowledge about the world written in cause-to-effect form to learn. The inputs and outputs of the learning system, but not its goals, are similar to those of an abductive reasoning system. Via a short case study, it was argued that this is an appropriate and natural way of representing knowledge for an inductive learner.

Acknowledgements

This paper benefitted greatly from discussions with many friends and colleagues, and from advice and encouragement from my advisor, Alex Borgida. Most of this research was done while the author was receiving a Marion Johnson Fellowship. The author is currently receiving an AT&T Fellowship.

References

[1] William W. Cohen. Abductive explanation based learning: A solution to the multiple explanation problem. Technical Report ML-TR-26, Rutgers University, 1989.

[2] Gerald DeJong and Raymond Mooney. EBL: An alternative view. *Machine Learning Journal*, 1(2), 1986.

[3] Tom Fawcett. Learning from plausible explanations. In *Proceedings of the Sixth International Workshop on Machine Learning*. Morgan Kaufmann, 1989.

[4] Haym Hirsh. Combining empirical and analytic learning with version spaces. In *Proceedings of the Sixth International Workshop on Machine Learning*. Morgan Kaufmann, 1989.

[5] Tom Mitchell, Richard Keller, and Smadar Kedar-Cabelli. Explanation-based generalization: A unifying view. *Machine Learning*, 1(1), 1986.

[6] Michael Pazzani. Selecting the best explanation in explanation-based learning. In *Proceedings of the 1988 Spring Symposium on EBL*. AAAI, 1988.

[7] Alfred Sheinwold. *5 Weeks to Winning Bridge*. Simon & Schuster, 1964.

[8] Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11), November 1984.

PED: A Technique for Refining Incomplete Determination-Based Theories

Sridhar Mahadevan
IBM T.J. Watson Research Center
P.O. Box 704, Yorktown Heights
NY 10598; Net: sridhar@ibm.com

Abstract

A major limitation of explanation-based learning (EBL) is that the domain theory used to explain training instances must be complete. This problem has been termed the *incomplete theory problem* [Mitchell *et al.*, 1986]. In this paper we present PED, a technique that extends EBL to incomplete theories containing determinations, a type of incomplete knowledge [Davies and Russell, 1987]. The key idea underlying PED is that training examples of a concept can be used to fill in gaps in a domain theory by propagating the information that they satisfy the target concept. Comparing PED to abduction-based techniques, such as LFP [Wirth, 1988], reveals two constraints on the gap-filling process that make PED more tractable than abduction-based techniques: one, only abduce predicates in the RHS of a determination; and two, constrain the search for relevant predicates to those in the LHS of a determination.

1 Introduction

A major limitation of explanation-based learning (EBL) is that the domain theory used to explain training instances must be complete. This problem has been termed the *incomplete theory problem* [Mitchell *et al.*, 1986]. In this paper we present one approach to the incomplete theory problem based on extending EBL to domain theories containing determinations, a type of incomplete knowledge proposed by Davies and Russell [Davies and Russell, 1987, Russell, 1986].

In particular, we describe PED, a techique that extends PROLOG-based EBL implementations, e.g. PROLEARN [Prieditis and Mostow, 1987], to determination-based theories. The key idea underlying PED is that training examples can be used to fill gaps in the domain theory by propagating the information that they are instances of the target concept; in contrast, EBL uses training instances to focus the explanation process and compute particular operational descriptions of the target concept.

Comparing PED to abduction-based techniques, such as LFP [Wirth, 1988], reveals two constraints on the gap-filling process that make PED more tractable than abduction-based techniques. One, PED only abduces predicates in the RHS of a determination. Two, PED constrains the search for relevant predicates to those in the LHS of a determination.

2 Determinations

Intuitively, determinations try to capture the notion of *relevance*. We say that an attribute P is relevant to an attribute Q if knowing that P holds for some object tells us something about whether Q holds for that object. A more precise definition is as follows:[1]

Definition 1 *Let P(x,y) and Q(x,z) be any two first-order sentences, where x represents the set of variables that occur free in both P and Q, while y and z represent the set of free variables that occur only in P and Q respectively. We say P(x,y) totally determines Q(x,z), or* $P(x,y) \succ Q(x,z)$, *iff*

$$\forall y, z[[\exists x P(x,y) \wedge Q(x,z)] \Rightarrow \forall x[P(x,y) \Rightarrow Q(x,z)]]$$

For example, let $P(x,y)$ denote the predicate $Nationality(x,y)$, meaning that individual x has nationality y. Also let $Q(x,z)$ denote the predicate $Language(x,z)$, meaning that x speaks language z. Then, the above definition states that if there exists an individual x whose nationality is y, and who speaks a language z, then all individuals of nationality y speak language z.

2.1 Determinations as a Form of Incomplete Knowledge

An example will help illustrate how determinations can be viewed as a form of incomplete knowledge. From

$$Nationality(x,y) \succ Language(x,z)$$

[1] See [Russell, 1986] for a description of other types of total determination.

and

$$Nationality(John, Us) \wedge Language(John, English)$$

it follows that

$$(\forall x) Nationality(x, Us) \Rightarrow Language(x, English)$$

However, just knowing that nationality determines language is not sufficient to compute an individual's language from his nationality. Examples are required to fill in this knowledge.

In general, from

$$P(x, y) \succ Q(x, z)$$

and

$$P(A, B) \wedge Q(A, C)$$

the implication

$$\forall x P(x, B) \rightarrow Q(x, C)$$

follows. This inference is a form of *single instance generalization*. We will make extensive use of this inference step later in the paper.

3 One View of the Incomplete Theory Problem

We begin by presenting one view of the incomplete theory problem, which is based on a discussion in [Russell, 1987]. Rajamoney and Dejong [Rajamoney and DeJong, 1987] proposed a classification of imperfect theory problems in EBL. In their terminology, the problem being studied here is the "broken explanations" problem. In other words, a complete explanation cannot be given because of missing rules in the domain theory. The missing rules manifest themselves as broken links in the explanation tree.

A central assumption in our approach is that the gaps in the domain theory are specifiable as total determinations. The domain theory is incomplete because there is insufficient information to evaluate queries using the determinations. Examples are needed to refine the determination into a set of implicative rules. It is this process of refinement that we study in this paper.

Figure 1 illustrates the general structure of the refinement process. Determinations may occur somewhere in the middle of an explanation path leading from the target concept to the training example description. The idea is to propagate the fact that the training instance satisfies the target concept, and show that the predicate P in the RHS of a determination holds. Similiarly, the predicates Q in the LHS of a determination are proven from the training instance description. Then, using the single instance generalization rule described above, a new implication can be added to the domain theory. (Pe and Ph are particular subsets of the domain theory.)

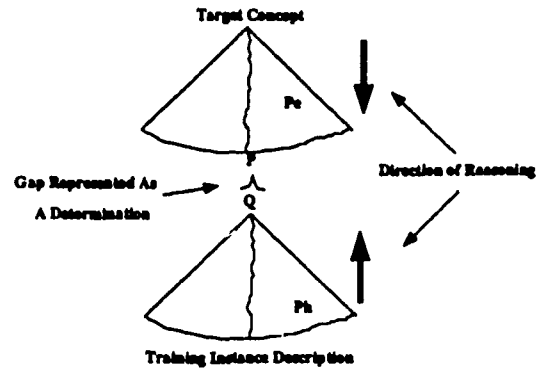


Figure 1: General Structure of Explanation in Determination-based Theories

4 PED: A Technique for Refining Incomplete Theories

This section describes PED, a technique for refining incomplete domain theories containing determinations. Figure 2 presents a high level description of PED. The top level procedure is called **assume**. Given the classification of the training instance `Q` (as an instance of the target concept), **assume** tries to explain `Q` from the training instance description. The procedure **explain** in Step 3 is basically EBL, implemented in techniques such as PROLEARN [Prieditis and Mostow, 1987]. The second argument `G` of **assume** represents a generalized (operational) sufficient condition of the target concept returned by PED.

The important steps are Step 2 and Step 4. Step 4 is invoked when EBL fails. Step 4 first retrieves a determination whose RHS unifies with `Q`. Next, it checks whether two of three conditions for single instance generalization hold: the query `Q` representing an instance of the RHS of the determination must be fully instantiated, and the instantiated LHS `P` (under the same variable bindings) of the determination must hold. At this point PED has located a possible gap in the domain theory, which if filled may allow the training example to be explained. The only remaining condition is that the query `Q` must hold. PED asserts `P` $\succ$ `Q` as a failed determination, and then backtracks trying other ways of showing that the example is an instance of the target concept.

Step 2 is invoked when the **explain** procedure fails to show that the training example is an instance of the target concept. PED first checks to see whether the cause of the failure was a failed determination, by retrieving the failed determination `P` $\succ$ `D` (which was stored in Step 4). PED now makes a type of closed world assumption: since all explanation paths except for the one using the determination `P` $\succ$ `D` failed, the RHS `D` must be true in order for the training example to be an instance of the target concept.

```
assume(Q,G) ←                          % Step 1
   explain(Q,G)

assume(Q,G) ←                          % Step 2
   Retrieve failed determination P ≻ D
   Explain the lhs P
   Use single instance generalization to
   create a new rule Dg ← Pg
   Assert the new rule in the knowledge base
   assume(Q,G).

explain(Q,G) ←                         % Step 3
% This step corresponds to standard EBL

explain(Q,G) ←                         % Step 4
   Retrieve P ≻ Q
   If the lhs P can be explained, and
      Q is ground
      Then assert P ≻ Q as a failed
      determination.  Backtrack and try
      other paths
```

Figure 2: The PED Procedure

PED next shows that P follows from the training instance. At this point, since instances of both the LHS and RHS of the determination have been shown, PED carries out the single instance generalization inference step, and creates a new implication. PED then adds the new rule to the domain theory, and recursively invokes the **assume** procedure. If this implication filled the only gap preventing the successful explanation of the training example, the second invocation of **explain** should terminate successfully, and return an operational sufficient condition for the target concept. Otherwise, further gaps in the domain theory may need to be filled using the above procedure.

4.1 Example

Consider the example domain theory shown in Figure 3 (this originally appeared in [Russell, 1987]). The target concept is `stack(X,Y)` meaning the class of pairs of objects that can be safely stacked on one another. The determination in the domain theory specifies that the material (`mat`) and construction (`constr`) of an object determine its fragility (`frag`). Suppose PED is given the query `assume(stack(box1,box2),G)`. The predicates `mat`, `constr`, and `wt` (meaning weight) are assumed to be operational.

First, the query `explain(stack(box1,box2),G)` is generated, which creates the goal `frag(box2,low)`. This goal fails because no information on box2's frag exists, and furthermore the determination for `frag` cannot be used analogically as there is no precedent in the

```
% target concept definition
stack(X,Y) ← frag(Y,low) ∨ lighter(X,Y).

% domain theory
mat(X,M) ∧ constr(X,C) ≻ frag(X,F).
lighter(X1,X2) ← wt(X1,W1), wt(X2,W2),
                 W1 < W2.
mat(X,Y) ← made_of(X,Y).
constr(X,Y) ← body(X,Y).

% training example description.
made_of(box1,lead).  made_of(box2,steel).
wt(box1,100).        wt(box2,10).
body(box1,rigid).    body(box2,rigid).
```

Figure 3: Example to Illustrate PED

knowledge base.[2] Since the instantiated LHS of the determination

`mat(box2,steel) ∧ constr(box2,rigid)`

can be shown from the training example, Step 4 declares the instantiated frag determination as a gap in the domain theory, which if filled could allow the example to be explained. PED then backtracks trying to prove `stack` through the other disjunct `lighter(box1,box2)`. This fails because `box1` is heavier than `box2`. Note that due to predicate completion, the sufficient condition for `lighter` is also a neccessary condition.[3]

At this point, Step 2 of PED is invoked. The failed goal `frag(box2,low)` is retrieved, which is now assumed true. Note that since the sufficient condition for `stack` is also a neccessary condition (due to completion), PED's reasoning at this point is of the form

"from $P \leftrightarrow Q \vee R$ and $\neg R$ and P, infer Q"

which is deductively valid. In the next part of Step 2, the LHS of the determination for `frag` is evaluated. The instantiated LHS and RHS of the determination are generalized to the rule

`mat(X,steel) ∧ constr(X,rigid) → frag(X,low)`

which is subsequently asserted. The procedure **assume** is invoked again on the original query. This time the call to **explain** succeeds, and PED finally returns with the result G =

`mat(Y,steel) ∧ constr(Y,rigid) → stack(X,Y)`

[2] Determinations can also be incorporated in a theorem-prover, such as PROLOG, as a form of analogy [Davies and Russell, 1987].

[3] PED uses predicate completion [Lloyd, 1984] to treat the disjunction of all the sufficient conditions of a predicate as a neccessary condition.

5 PED As A Constrained Abduction System

In this section we discuss the relation between abduction and PED, showing how PED can be viewed as doing a constrained form of abduction. We also compare PED to several techniques that are based on abduction, such as LFP [Wirth, 1988].

To see how PED can be viewed as performing a restricted version of abduction, we return to Example 1 above (see Figure 3). From

`stack(X,Y)` ← `lighter(X,Y)` ∨ `frag(Y,low)`

two explanations for the target concept instance `stack(box1,box2)` follow, namely `lighter(box1,box2)` and `frag(box2,low)`. PED abduces that the latter is the best explanation because it is defined using a determination, even though both hypotheses cannot be shown from the implicative portion of the domain theory. Therefore PED can be viewed as using the heuristic – *Given a choice, abduce predicates that appear in the RHS of a determination*[4] – to filter out the set of possible hypotheses that can explain a given fact.

Sometimes this heuristic is not sufficiently powerful since there may be several predicates that are defined using determinations. To deal with such situations, PED needs to be extended to use additional heuristics, such as simplicity, for selecting among competing predicates.

Comparing PED with Abduction-based Techniques We now compare PED to the growing number of learning techniques that are based on abduction, such as LFP[Wirth, 1988], a technique used in the ODYSSEUS system[Wilkins, 1987], CIGOL[Muggleton and Buntine, 1988], and a technique described by O'Rorke[O'Rorke, 1989]. There are two differences between PED and these other techniques: one, PED assumes that the gaps in the domain theory are filled by determinations; two, PED selectively abduces predicates that are defined using determinations. As a consequence of these two differences, PED suffers less from the combinatorial explosion of possible predicates that can be abduced, as well as the many possible ways in which rules can by hypothesized to fill the gaps. On the other hand, abduction-based techniques are more powerful than PED in filling in gaps in a domain theory since they do not require that the determinations be known.

To illustrate the point that abduction-based techniques are faced with a more serious combinatorial explosion problem, let us examine the LFP technique proposed by Ruediger Wirth[Wirth, 1988, Wirth, 1989]. LFP is based on Muggleton's idea of inverse resolution[Muggleton and Buntine, 1988]. LFP is similiar to PED in that it uses a training example to fill in gaps in an incomplete but correct Horn theory.[5]

The steps behind LFP are as follows. Given a training instance, LFP first tries to explain the training instance using the partial domain theory. If a particular predicate fails in the proof process, for example if it is not defined, LFP asks an oracle if that predicate is true. If it is true, then LFP proceeds to the next subgoal, otherwise it backtracks. Eventually, the first phase terminates in a *partial proof tree*. Leaves in this partial proof tree that were justified by the oracle are denoted by a special label, since these represent gaps in the domain theory that need to be filled.

To illustrate the operation of LFP, let us consider the domain theory in Figure 3 with the fragility determination deleted. At the end of phase 1, LFP will construct the partial proof tree

`safe_to_stack(box1,box2)` ← `fragility(box2,low)`

LFP will affix a special label to fragility predicate since the oracle was used to justify the truth of the predicate, and thereby bottom out the proof.

The second phase of LFP is constructing a *complete proof tree*. Since the domain theory is incomplete, LFP can only approximate the complete proof tree. Basically what LFF does is to examine the training instance description for facts that are relevant to labelled leaf predicates in the partial proof tree. The idea is to find some link between the labelled leaf nodes and facts in the training instance, which can be used to complete the partial proof tree. LFP uses heuristics to guide it in computing the relevant facts. Using Example 1 again, LFP might decide that `made_of(box2,steel)`, `weight(box2,10)`, and `body(box2,rigid)` are relevant because the constant symbol `box2` appears in them (and in the labelled fragility predicate also). (This is an actual heuristic used in LFP to compute relevance.) In contrast, PED uses determinations to compute the relevant facts. LFP finally hypothesizes the following complete proof tree:

`safe_to_stack(box1,box2)`
↑
`fragility(box2,low)`
↑
`made_of(box2,steel)` ∧ `weight(box2,10)` ∧ `body(box2,rigid)`

The third phase of LFP is to abstract the leaf nodes in the complete proof tree by forward chaining on the implicative rules in the domain theory. In particular, using the following implications in the domain theory

[4] I thank Thorne McCarty for this observation.

[5] In a subsequent paper[Wirth, 1989], Wirth describes an improved technique LFP2 that relies less on an oracle to construct partial proof trees, and which also can invent new terms using Muggleton's inverse resolution technique. It is more appropriate to compare here PED to LFP since it highlights better the main differences between PED and abduction-based techniques.

```
material(X,Y) ← made_of(X,Y)
construction(X,Y) ← body(X,Y)
```

LFP constructs the following modified complete proof tree

```
                safe_to_stack(box1,box2)
                          ↑
                fragility(box2,low)
                          ↑
material(box2,steel) ∧ weight(box2,10) ∧
        construction(box2,rigid)
```

LFP next compares the partial proof tree and the complete proof tree to try to hypothesize missing rules in the domain theory. For Example 1, it may hypothesize the following rule:

```
fragility(box2,low) ← material(box2,steel) ∧
                      weight(box2,10) ∧
                      construction(box2,rigid)
```

Finally, LFP generalizes the above rule using heuristics. In the description of LFP, Wirth uses heuristics that are particular to natural language parsing, which is the domain of application. Subsequently, in LFP2, Wirth uses inductive techniques such as maximally specific generalization to generalize from multiple examples of such instantiated rules.

The differences between LFP and PED should be clearer now. First, PED uses determinations to compute the relevant facts, whereas LFP uses heuristics. The heuristic used above can easily be fooled by many irrelevant facts. For example, the predicate `weight` may not be relevant to fragility. In fact, if the predicate `owns(box2,john)` was present in the training instance description, LFP would have included this in the body of the rule hypothesized to fill the gap in the domain theory. Second, PED uses a justified form of single instance generalization, whereas LFP uses inductive learning techniques whose effectiveness depend on the generalization language containing the right abstractions. Thirdly, all the phases of LFP could potentially lead to a combinatorial explosion of possibilities. For example, the third phase of LFP involves forwarding chaining on the rules in the domain theory which could lead to many alternative possibilities for abstracting the leaf nodes in the proof tree.

A detailed comparison of PED with the other abduction-based techniques cited above is given in [Mahadevan, 1990].

6 Conclusions

In this paper we described the PED technique, which extends EBL to incomplete determination-based domain theories. PED uses training examples to fill gaps in a domain theory by propagating the information that they satisfy the target concept definition. Gaps in the domain theory are specified using determinations. PED fills the gaps by extracting implicative rules from determinations. An example was presented that illustrated PED's ability in refining the determinations in a domain theory. Finally, PED was compared to abduction-based techniques.

References

[Davies and Russell, 1987] T. Davies and S. Russell. A logical approach to reasoning by analogy. In *IJCAI*. Morgan Kaufmann, 1987.

[Lloyd, 1984] J. W. Lloyd. *Foundations of Logic Programming*. Springer-Verlag, 1984.

[Mahadevan, 1990] S. Mahadevan. *An Apprentice-based Approach to Learning Problem-Solving Knowledge*. PhD thesis, Rutgers University, 1990.

[Mitchell *et al.*, 1986] T. Mitchell, R. Keller, and S. Kedar-Cabelli. Explanation-based generalization: A unifying view. *Machine Learning*, 1(1), 1986.

[Muggleton and Buntine, 1988] S. Muggleton and W. Buntine. Machine invention of first-order predicates by inverting resolution. In *Proceedings of the Fifth IML Conference*. Morgan-Kaufmann, 1988.

[O'Rorke, 1989] P. O'Rorke. Theory formation by abduction: Initial results of a case study based on the chemical revolution. In *Proceedings of the Sixth Machine Learning Workshop*. Morgan Kaufmann, 1989.

[Prieditis and Mostow, 1987] A. Prieditis and J. Mostow. Towards a PROLOG interpreter that learns. In *Proceedings of the Sixth AAAI*. Morgan-Kaufmann, 1987.

[Rajamoney and DeJong, 1987] S. Rajamoney and G. DeJong. The classification, detection, and handling of imperfect theory problems. In *IJCAI*. Morgan Kaufmann, 1987.

[Russell, 1986] S. Russell. *Analogy and Inductive Reasoning*. PhD thesis, Stanford University., 1986.

[Russell, 1987] S. Russell. Analogy and single instance generalization. In *Proceedings of the Fourth IML Conference*. Morgan-Kaufmann, 1987.

[Wilkins, 1987] D. Wilkins. Knowledge base refinement using apprenticeship learning techniques. In *IJCAI*. Morgan Kaufmann, 1987.

[Wirth, 1988] R. Wirth. Learning by failure to prove. In *3rd European Working Session on Learning*. Pitman, 1988.

[Wirth, 1989] R. Wirth. Completing logic programs by inverse resolution. In *4th European Working Session on Learning*, 1989.

Plausible Inference vs. Abduction

Gerald DeJong

Computer Science Dept. & Beckman Institute
University of Illinois

Introduction

The process of constructing explanations is particularly relevant to Explanation-Based Learning researchers. It has become clear that the notion of explanations as a deductive proof as manifested in EBG [Mitchell86] and EGGS [Mooney86] is limited in scope and constrains EBL systems to relatively sterile micro-world domains. The number of things that can be proved is small and the set contains little of interest that could not be inexpensively constructed from first principles as needed. Something more than truth-entailment inference is needed to drive explanation construction.

When confronted with experiences from the real world (or even an artificial but rich domain) EBL systems react with brittleness. In particular, world experiences often include observations which directly contradict deductive conclusions of the system. This results in the system entering some kind of internal should-not-occur state from which conventional EBL systems cannot recover. Let us (somewhat generously) define a non-brittle EBL system:

A non-brittle EBL system is one which tolerates any observation or set of observations of the real world.

Of course, if a system is intentionally or unintentionally given contradictory experiences, it may reach a should-not-occur state. The observations of such experiences could not have been of the real world since the real world is always self-consistent. A non-brittle system need not tolerate *any* input; only those that are faithful to the real world; as long as we do not lie to it by faking world observations, its code must not hang.

Now consider what such an EBL system's explanation structures might look like. Logical proofs must be rejected as explanations because of the *qualification problem*, which is intuitively well stated (though a bit circular) in [Genesereth87]:

> Most universally quantified statements will have to include an infinite number of qualifications if they are to be interpreted as accurate statements about the world.

The problem was introduced by McCarthy as an aspect of the frame problem [McCarthy69] and has been discussed much. As an example, consider the classic implication about birds flying:

$$\forall\, x\, [\text{Bird}(x) \Rightarrow \text{Flies}(x)]$$

This FOPC sentence overstates the case for flying because, after all, not all birds fly. In particular, penguins do not fly so to be faithful to the world the rule must be amended:

$$\forall\, x\, \{[\text{Bird}(x) \wedge \neg\, \text{Penguin}\, (x] \Rightarrow \text{Flies}(x)\}$$

But, of course, ostriches cannot fly, nor can emu's, nor can kiwi's, nor can many other increasingly exotic birds. The rule could be patched for these, but there are other problems. A bird with a broken wing cannot fly so the rule must again be amended. A dove missing more than five flight feathers on one side cannot fly, nor can an eagle missing more than 12, nor can... Once again the rule must be fixed. But, of course, we are not finished. A bird that has been appropriately conditioned in a Skinner box cannot fly, a bird that has been cooked for dinner cannot fly, a bird that is attached to an anvil cannot fly, etc. The list is endless. This problem is not specific to birds and flying. As pointed out by Genesereth and Nilsson it applies to nearly all universally quantified sentences intended to describe the real world.

If we are to allow our system's domain knowledge to specify general statements about the world (i.e.,

universally quantified sentences) those statements must necessarily entail some conclusions which are contradicted by reality. Thus, a logically sound inference procedure like resolution or backward chaining through horn clauses, will necessarily violate the non-brittleness definition.

To achieve non-brittle EBL systems the process of constructing an explanation must take on more the flavor of imposing an interpretation on an example and less of theorem proving. The veracity of the explanation can no longer guaranteed, nor can the "generalization" of such an explanation be defined via strict inclusions of possible world states. To allow such inferences the force of logical entailment invites contradictions. It is highly desirable that the EBL process not be truth preserving. This statement would surprise many. However, the complexity of real-world situations and the impossibility of engineering a complete and correct domain theory dictates it.

Unburdened by the thorny crown of truth-preserving inference, EBL systems must substitute some other mechanism to take advantage of existing background knowledge.

Abduction

The obvious possibility is abductive inference. Abduction has long been associated with the notion of "explanation". Furthermore, abductive inference, as commonly construed, is not sound in the formal logic sense. Since explanation-based learning systems must support their explanations through some formalism, and since EBL systems seem to benefit from some kind of unsoundness, abduction may serve well as a formal basis for explanation-based learning.

The standard interpretation of abduction is as an inference rule of the form:

$$\frac{A \Rightarrow B \qquad B}{A}$$

That is, from knowing an implication and its consequent, hypothesize that its antecedent holds.

A straightforward use of abduction in EBL might identify B with the goal concept, $A \Rightarrow B$ as an element of the the complete and correct background knowledge, and A as the operational sufficient condition. Abduction offers a way to conjecture A as the explanation of B.

While abduction has many desirable properties, I believe it is deficient as an underlying formalism for explanation construction in explanation-based learning. While abduction is not truth preserving, it is still too strong an inference formalism. The reason can be seen as an interaction between the non-brittleness definition and the qualification problem.

Let us consider the source of abduction's unsoundness. Informally, we can see that the above abduction rule suggests that B, which is observed in the training example, came about because of A. The rule is unsound because there there is the possibility of other implication rules in the system that may have lead to B. Our knowledge might, for example, have included the sentence $C \Rightarrow B$. Clearly, then, C is as good an explanation for B as A. Without ruling out the possibility of a $C \Rightarrow B$ rule being responsible for the truth of B, the inference of A from B and $A \Rightarrow B$ is unsound. The problem is that there exists enough information in the axiom set to make the inference sound.

A sentence describing the conditions for the sound inference of A as the explanation of B is already entailed by the background knowledge. We simply include sufficient antecedents in the implication to insure that no other possible cause of B can apply. For example, if $C \Rightarrow B$ and $A \Rightarrow B$ were the only ways to infer B, the resulting entailed sentence would be $B \wedge \neg C \Rightarrow A$. A finite such sentence can be constructed for any explanation. Thus, the qualification problem does not arise and we can be sure that our system no longer applies to the real world.

This works because there is a necessary and sufficient specification of B entailed by the domain theory. The above argument is valid only if we have closed the world on our domain theory. If there is no way to collect all sentences of the form $\Psi \Rightarrow B$, then the new sentence cannot be constructed. So perhaps the argument is not against abduction so much as against closing the world of our domain theory. An unsound inference (say abduction) along with a domain theory that is never assumed to be closed, neither runs afoul of the qualification problem nor violates the non-brittleness principle. But such a system only superficially has the right properties and it has them for all the wrong reasons. If a world observation is inconsistent with a conclusion of the system, the system simply takes it back; none of its conclusions are particularly believable since it has an unsound inferencer. If a world observation is outside the scope of its theory it can be blamed on its admittedly incomplete model. It avoids brittleness and the qualification problem by not having many opinions and not believing strongly in the ones it has.

Plausible Inference

Instead, I believe that some form of "plausible" knowledge and inference is needed, which must necessarily be rather different from abduction, at least in its normal guise. Semantics will be altered to weaken domain theory statements rather than to compensate for their inaccuracies through incompleteness and inferential unsoundness. Unlike [Collins86], however, the motivation of this form of plausible inference is entirely computational adequacy. No psychological claims or justifications are being advanced.

In a theory of plausible inference an explanation is an educated, somewhat abstract guess at why the proposition is likely to be true given what is believed. For example, one might plausibly reason that since it is autumn in Central Illinois, tomorrow will be a windy day. This illustrates the two hallmarks of our plausible inferences: First, they are not certain. It is entirely possible that tomorrow will not, in fact, be windy in Central Illinois. Second, plausible inferences are often abstract. It is not plausible to conclude that the winds will be out of the north northwest at 22 mph. To be an acceptable rule the characterization of the wind must be much more abstract.

I propose an approach to plausible inference where implication has a different semantics. I will continue to write sentences like:

$A \Rightarrow B$

But by this I mean

$\Phi \wedge A \Rightarrow B \wedge \Psi$

using the standard semantics for implication.

There may be conditions under which "A" is satisfied but "B" is not true. Φ represents a specification of the context in which the plausible rule is guaranteed. Φ specifies the implicit assumptions built into the plausible rule "$A \Rightarrow B$". Ψ, on the other hand, specifies those things in the world that are guaranteed even though there is no explicit way to conclude them from the plausible implication.

To be a useful rule to the plausible inference system, the conditions that make Φ false should be, for the most part, infrequent or otherwise uninteresting.

Much of the power of this approach is traceable to the fact that no attempt is made to specify the context conditions of a domain rule (such as "Φ" in the above implication), while acknowledging the possibility that they may not be met. Such context conditions must not be represented or directly reasoned about.

In this view of plausible inference, the requisite unsoundness is removed from the inference rule and embedded in the world knowledge itself. Thus, modus ponens and other sound inference mechanisms can be used. This is, in a sense, the dual of the abduction approach in which unsoundness is introduced directly by the inference rule. The advantage for EBL is that the unsoundness of a conclusion is a function of the knowledge used in the explanation which is declaratively specified in the explanation, while the inference rules (modus tolens, resolution, abduction...) are implicit. Schemata (or macro–operators) generalised from explanations can thus be independently evaluated for their adequacy in the real world. For some applications, at least, this property supports a kind of convergence for the learning that would be difficult or impossible using abduction.

An Example

This notion of plausible inferencing has been implemented in an EBL system that learns to plan in continuous domains. Its primary domain is that of controlling the speed of a single gear automobile by manipulating the gas and clutch controls. The domain knowledge is in the form of plausible qualitative proportionalities among quantities. For example, there is a quantity that represents the current position of the gas control. Call it GAS-PEDAL-POSITION. There is another quantity that represents the rate of GAS-FLOW. One domain theory rule specifies that these are qualitatively positively proportional:

INCREASE(GAS-PEDAL-POSITION, *interval*) $\Rightarrow$
INCREASE(GAS-FLOW, *interval*)

and

DECREASE(GAS-PEDAL-POSITION, *interval*) $\Rightarrow$
DECREASE(GAS-FLOW, *interval*)

which means that in some implicit context the flow of fuel can be increased by advancing the throttle. This is not always the case – the tank may be empty, the fuel line blocked, etc.

The system is given the goal concept of accelerating the car from 0 to 30mph. It allowed to observe a training episode in which an expert solves the problem by manipulating various controls (including the air conditioner temperature, the throttle, the car radio, and the clutch). The system pieces together a plausible explanation for the expert's actions from its domain theory. Chaining plausible implication rules together yields a valid conclusion only when the rules' implicit contexts overlap with each other and with the real world situation. This overlap can never be confirmed, but it can be denied if a world observation contradicts the conclusion. The system first assembles a plausible explanation for the car going faster tracing the car's velocity

through the engine rpms, gas flow, and throttle position; it does not include radio, a/c, or clutch controls. There exist contexts in which this explanation corresponds to reality (e.g., a/c off, radio volume and clutch set to median values). However, in the next planning problem the system does not stay within this context, the conclusion is rejected, and the next most plausible explanation is constructed and generalised. After two more tries the system constructs an explanation that is sufficient to control the car's speed. As it happens the working explanation is not completely correct either. It implicitly assumes, among other things, small accelerations and no hills. However, the profile of planning problems given to the system never violates these constraints and with experience the system continues to become more accurate and smooth in its velocity changes. See [DeJong89] for details.

Conclusions

I suggest that the problem with explanation generation for EBL systems is not with the inferencer but with the semantics of the domain knowledge itself. Domain rules must not overstate their knowledge with regard to the qualification problem. This can be overcome by adopting a *plausibility* semantics for these rules. Each rule is logically valid only within an implicit context. Importantly this context can never be known, represented, or reasoned about. Conclusions inferred from such rules are not "right" or "wrong" but rather they have their own derived implicit context which any world situation may or may not satisfy. Experience and feedback from the world is essential. It is only through such world observations that the system can discover the mismatch to a conclusion's implicit context. EBL-acquired schemata provide a convenient memory hook to store combined analytical/experiential planning knowledge. Finally, while the discussion has been couched in terms of EBL the plausible inference approach may be useful for other aspect of automated reasoning.

Acknowledgement

This research was supported in part by the Office of Naval Research under grant N-00014-86-K-0309.

References

[Collins86] A. Collins and R. Michalski, "The Logic of Plausible Reasoning: A Core Theory," UILU-ENG-86-1775, ISG, University of Illinois at Urbana-Champaign, 1986.

[DeJong89] G. DeJong, "Explanation-Based Learning with Plausible Inferencing," *Proceedings of the 1989 European Workshop on Machine Learning*, Montpellier, France, December, 1989, pp. 1-10.

[Genesereth87] M. Genesereth and N. Nilsson, *Logical Foundations of Artificial Intelligence*, Morgan Kaufmann, Palo Alto, CA, 1987.

[McCarthy69] J. McCarthy and P. J. Hayes, "Some Philosophical Problems from the Standpoint of Artificial Intelligence," in *Machine Intelligence 4*, B. Meltzer and D. Michie (ed.), Edinburgh University Press, Edinburgh, Scotland, 1969.

[Mitchell86] T. M. Mitchell, R. Keller and S. Kedar-Cabelli, "Explanation-Based Generalization: A Unifying View," *Machine Learning 1*, 1 (January 1986), pp. 47-80.

[Mooney86] R. J. Mooney and S. W. Bennett, "A Domain Independent Explanation-Based Generalizer," *Proceedings of the National Conference on Artificial Intelligence*, Philadelphia, PA, August 1986, pp. 551-555. (Also appears as Technical Report UILU-ENG-86-2216, AI Research Group, Coordinated Science Laboratory, University of Illinois at Urbana-Champaign.)

The Use and Evaluation of Contextual Knowledge for Explanation Completion

Andrea Pohoreckyj Danyluk

Columbia University - Department of Computer Science
New York, New York 10027
(212) 854 - 8121
email: andrea@cs.columbia.edu

Introduction

The use of deduction as a mechanism for generating good explanations is appealing. It is truth-preserving and can be performed on an example-by-example basis. The problem with deduction is the reliance upon a theory, or base of laws, from which inferences are made. In many - or, more realistically, most - domains, it is not safe to assume that such a theory will be perfect. In the explanation-based learning (EBL) community, which uses deduction as a basis for its method, this is now being addressed as a major problem. (See, for example, [Rajamoney and DeJong 87] for a discussion of imperfect theory problems in EBL. The relationship between EBL and abduction was first described in [O'Rorke 88].) An alternative to deduction is to use techniques such as statistical methods that require less in the way of built-in application-specific theories. These, however, generally require large example bases, which are not always readily available.

In this paper, we take abduction to be inference to the best explanation. Given this, we make two claims here. The first is that we believe an effective mechanism for abduction can be found by combining elements of both deduction and induction. Treating inductive methods as a necessity stems from the belief that the domain theories available for deductive inference will often be imperfect. There are three major ways in which a theory may be faulty: it may be incomplete; it may be incorrect; or it might be intractable to use. This paper does not assume that the theory used by deductive inference is perfect. However, it concentrates solely on incomplete theories. That is, we assume that correct partial explanations can be generated. Although falling back on induction as a realistic necessity for completing partial explanations, much knowledge can be brought to bear from the partial explanations as well as from the deductive process in general. We refer to this as *contextual knowledge*. Such information may be used to provide focus on past examples in order to form a set from which the knowledge missing from a domain theory can be learned and then instantiated to complete a partial explanation. Additionally, it may focus attention away from attributes of that learned knowledge in order to make it more generally - yet still correctly - applicable later.

This paper is organized as follows. Section 2 introduces different types of contextual knowledge that might be brought to bear in completing partial explanations generated by deduction. Section 3 describes a mechanism for evaluating the effectiveness of various types of contextual knowledge as well as the explanations derived using them. In Section 4 we describe work completed to date. We conclude with a discussion of our further goals in the investigation of context.

Contextual Knowledge: Identification and Application

Consider explanation construction in the domain of network fault diagnosis, specifically an ethernet/token-ring network. In this domain, a fault is signalled by the inability of one user to reach another connected to the same network. The knowledge used by the performance system is encoded on a level that enables isolation of a fault to a particular segment of a network, but does not allow deeper analysis of the problem. Given an input pair consisting of a diagnosis as well as a frame description of the network's state at the time of diagnosis, an explanation might be constructed linking particular features of the state description to the diagnosis, as in Figure 1. The explanation is drawn as a proof tree with the diagnosis, or goal, at its root. Leaves refer to features of the input state description. Abbreviations are used, as the tree shown is system output; the tree may be read from the bottom as follows:

> If the target of an incomplete communication responds in general, while the user initiating the incomplete connection cannot seem to reach anything in the network, then the problem appears to lie with the source of the incomplete connection (rather than with the destination). If the problem appears to lie with the

source, and the user at the source is on a token ring network, then etc.

The explanation in Figure 1 is constructed deductively by backward chaining from the goal to the input state description using a domain theory for network fault diagnosis. A problem would occur if any of the information in the domain theory were missing. An example of this type of problem is shown in Figure 2. In order to complete such a partial explanation one might apply an inductive method that would consider other examples of the concept *LOOKS-LIKE-SOURCE-PROB* in order to find those features that correctly imply it. Alternatively, one can look more closely at the domain theory as well as the partial explanation derived. Say this domain theory is constructed in such a way that input features are never referred to more than one time in any given explanation. Then in determining those that best imply the uninferred subgoal, one can ignore all those features already used. This is an example of the use of contextual knowledge when completing an explanation. In general, contextual knowledge includes:

- Attributes of **inputs** - for example, features and their values, combinations of features, etc.
- Attributes of both partial, and earlier complete, **explanations** - for example, the specific explanation goal, parts of the domain theory used in constructing the explanation, structure of the explanation (i.e., shape of the proof tree), etc.
- Attributes of the **domain theory** - for example, the origin of the theory, etc.
- Attributes of the **history** of the explanation system - for example, the number of explanations constructed, the content of explanations constructed, relationship of past complete explanations to a current one, etc.

In general, this information can be used by either focusing attention toward or away from specific attributes.

Criteria for Evaluating the Application of Contextual Knowledge

There are many possible sources of contextual knowledge, as described above. Clearly, not all of this is necessarily useful. Two potential issues for concern are correct and efficient application of contextual knowledge. We ideally want to use this additional information to generate correct explanations. We also want to do so in an efficient manner. There is no reason to expend resources to use knowledge whose application, while not incorrect, does not provide any additional information. This section addresses the issue of using contextual knowledge appropriately.

Certain types of information about the structure of contextual knowledge can guarantee the correctness of applying it. For example, if we know that a domain theory is structured in such a way that individual input features are never used multiple times within a single explanation,

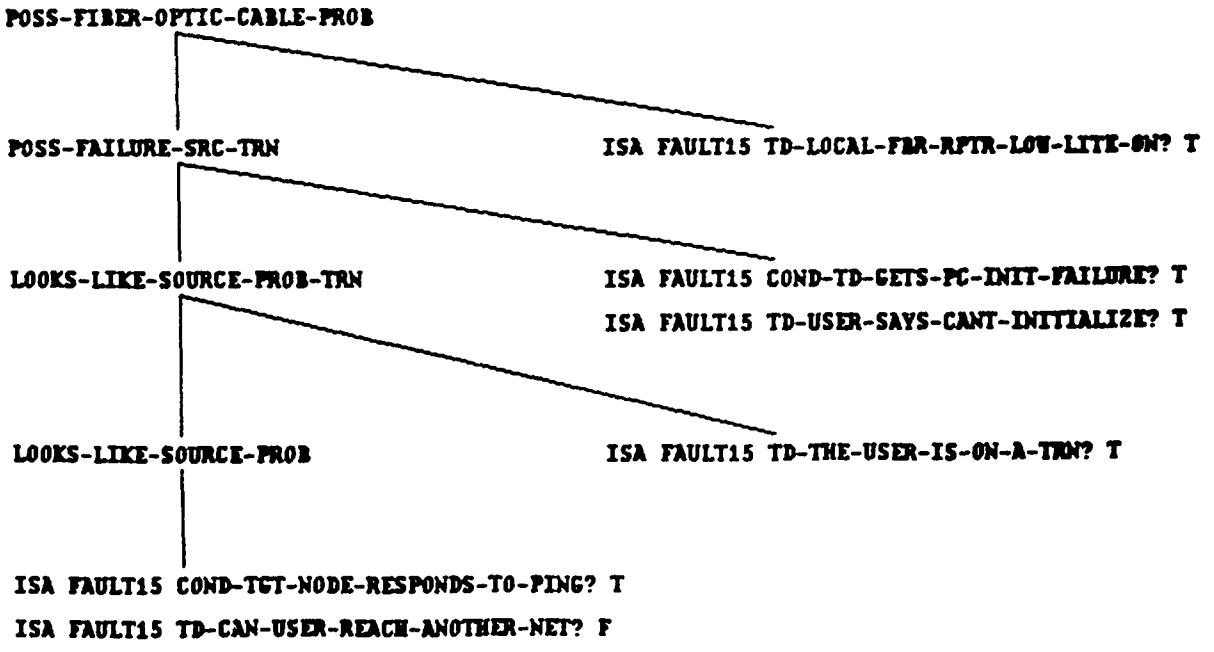


Figure 1: An Explanation for Network Fault Diagnosis

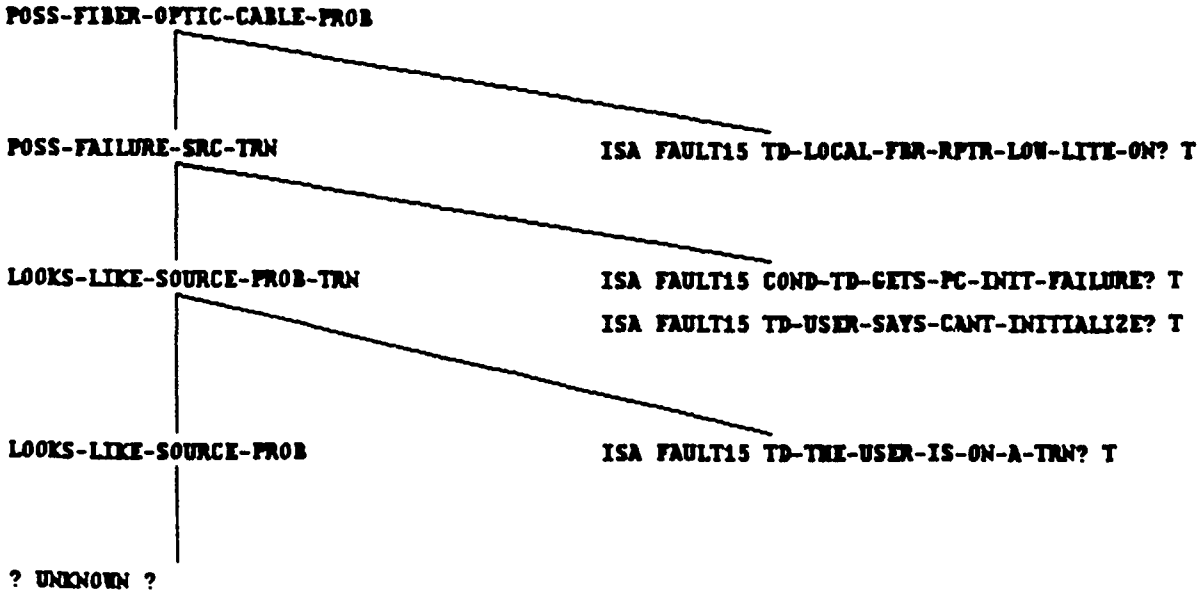


Figure 2: A Partial Explanation for Network Fault Diagnosis

then we can prove that those already used in an incomplete explanation will not play a role in the part that is missing. This, of course, requires that we make assumptions about the correctness of the partial explanation derived. In general, we cannot expect assumptions of this type to hold true. Nor can we expect to have a complete theory of the contextual knowledge for any particular domain. Therefore, we are concentrating on empirically characterizing the performance of various types of contextual knowledge.

In order to clarify the scope of our investigation, we make a number of assumptions. The first is that our domain theory, represented as a rule base, does not contain any incorrect rules, as indicated in the introduction. We also assume that the domain theory is tractable. The nature of the missing knowledge is that entire rules are missing. In our representation, this means that there is either no way to deduce a given subgoal with the partial domain theory, or that there are disjunctive ways but one or more of the disjuncts is missing, thus not covering all the cases in which a subgoal should be deducible. We assume that the system receives no noisy input. We define noisy input to be any pair of a goal and an example, where the goal would not be deducible from the example using *any* correct theory.

In order to generally characterize the applicability of contextual knowledge, we are investigating a number of domains. To date we have looked at network fault diagnosis, radio fault diagnosis, and terrorist event news stories. For each domain, we begin with a complete, correct, and tractable domain theory. We delete rules from the theory in order to answer the following questions:

1. Can the missing rules be created using contextual knowledge in order to complete partial explanations that would have been complete had the rules not been deleted? In order to evaluate the relative effectiveness of various types of context knowledge we must evaluate the explanations derived using them. The criterion we are using is the closeness of the explanation to the one that would have been generated deductively had the theory been complete. Closeness of the derived explanations to those specified by the original theory provides at least one reasonable measure of "goodness", assuming that the domain theory was designed by an expert.
2. If created rules differ from those deleted, to what extent do they differ? Ideally, if efficiency is one of our goals, we would like to find not just the correct instantiation of a particular rule to complete a partial explanation, but its correct generalization, so that it can be used later. That is, we would like to learn rules for later use.
3. What combinations of contextual knowledge appear to lead to better rules most quickly?

We are performing extensive tests varying parameters corresponding to the selected domain, the degree of completeness of the domain theory, and subsets of the various possible contextual types. All tests are being performed within a system developed by us, called Gemini. Gemini is described in detail in [Danyluk 89a].

Investigation of the Effectiveness of Using Context Information: Results to Date

In this section we discuss some of the results of our investigation into the application of contextual knowledge to explanation completion. Specifically, we describe some of the test runs performed with Gemini. All runs described in this section were done in the domains of network fault diagnosis and radio fault diagnosis. The complete rule base for the network fault diagnosis domain described above contains 56 rules. It was encoded from a prototype knowledge base that was extracted from experts maintaining the CMU campus computer network [Eshelman 88].

The radio fault diagnosis domain is similar to the network domain in that the theory is encoded on a level that enables isolation of a fault to a particular major radio component, but does not allow deeper analysis of the problem. The specific radio is a military communications radio. The rule base, containing 35 rules, was encoded from troubleshooting charts published in the operations manual for the radio [Radio Manual 86]. Input frames for this domain contain 21 slots.

Test I was performed in the domain of network fault diagnosis. For this test a single rule was removed from the complete domain theory. Instantiation of this rule in a proof tree would always place it at the leaves. In this test, three separate sets of contexts were studied. The first set used very little contextual information: to complete the explanation it selected common features from past examples that were most similar to the example being explained. This is essentially an implementation of similarity-based learning (SBL). The second set used additional context information that removed all features already appearing in the partial explanation. The third set additionally used a third type of contextual knowledge: it removed from consideration input features found consistently in examples throughout the history of the system's operation. The number of past examples considered was varied from 2 to 10. We found that no set of context knowledge produced incorrect results (i.e., either explanations or rules), although the rules created in the second and third sets were more general, with the third set giving best results. As expected, the results were better for each test set as more past examples were considered. Results were averaged over ten runs.

Test II was almost identical to Test I, except for the mechanism used to retrieve past examples. A less conservative approach was used where examples were selected randomly from a set of examples considered to be similar to the current one within a specified threshold. Results here were generally better than those of Test I for all test sets run. Results were averaged over two runs.

Test III was yet another variation on Tests I and II, but here retrieval was done randomly. Results were, understandably, most varied. They ranged from being general and correct to over-general, and therefore incorrect.

In **Test IV** we used the same three context sets as for Test I, but instead applied them to a case where the gap in the partial explanation would be essentially in the middle of a proof tree, rather than at its leaves. Again, results were good, giving only correct explanations and rules. However, the rules were judged to be less generally applicable than those found for Test I.

Test V was performed in the radio fault domain, using the same parameters as in Test I. Results here were similar to Test I in that better results were obtained when more contextual knowledge was used. With as few as three examples being retrieved from the system's memory, however, the explanations derived using the largest context set were *identical* to those that would have been found by the initial complete theory. That is, at that point essentially "perfect" explanations and rules were being found.

The results of the test runs may be summarized briefly as follows. Applying explicit contextual knowledge can indeed result in better rules (and thus explanations) than can be found by an inductive method alone. They act to significantly reduce the number of examples that must be considered. Furthermore, the selection of examples - contextual information in its own right - has potentially significant impact on the generality of the learned rule. Examples too similar to each other leave less room for generalization, while selecting examples too different may result in incorrect generalization. Finally, although it is possible to use contextual information to complete explanations with gaps in the middle, the new explanation will tend to be less good than when the gap is at the leaves. This occurs because less information is available from the partial explanation. These, as well as other tests, are discussed in detail in [Danyluk 89b].

Further Work in the Investigation of Context

In this paper we have discussed the use of contextual information to complete partial explanations that have been derived deductively. Contextual information is varied, however, and is not necessarily correct or useful to consider in all cases. We have performed a number of tests in order to determine the relative effectiveness of some different types of contextual knowledge. A great deal of investigation remains to be done before we have a clear understanding of the role each type of knowledge plays. As a step toward a more complete characterization of the effectiveness of using context knowledge, we are in the process of performing more complete and varied tests

using Gemini. These include more tests in the domain of terrorist event news stories as well as an extended version of the radio fault domain. We have recently formalized our notion of contexts and their use so that their combinations may be systematically generated. This will assure us greater coverage in testing.

Acknowledgments

This research was supported in part by the Defense Advanced Research Projects Agency under contract N00039-84-C-0165. Thanks go to Michael Lebowitz for his helpful comments on an earlier draft of this paper.

References

[Danyluk 89a] Danyluk, A. P. Finding New Rules for Incomplete Theories: Induction with Explicit Biases in Varying Contexts. Technical Report CUCS-466-89, Columbia University Department of Computer Science, 1989.

[Danyluk 89b] Danyluk, A. P. Recent Results in the Use of Context for Learning New Rules. Technical Report TR-89-066, Philips Laboratories, 1989.

[Eshelman 88] Eshelman, L. . . Personal Communication

[O'Rorke 88] O'Rorke, P. Automated Abduction and Machine Learning. Proceedings of the AAAI Symposium on Explanation-Based Learning, Stanford University, 1988, pp. 170 - 174.

[Radio Manual 86] Department of the Army. Technical Manual - Unit Maintenance. Technical Report TM 11-5820-890-20-1, Department of the Army, 1986.

[Rajamoney and DeJong 87] Rajamoney, S. and DeJong, G. The Classification, Detection and Handling of Imperfect Theory Problems. Proceedings of the Tenth International Joint Conference on Artificial Intelligence, Milan, Italy, 1987, pp. 205 - 207.

Constructing and Refining Non-monotonic Plan Explanations for Explanation-based Learning

Steve Chien

Beckman Institute, University of Illinois

405 N. Mathews Avenue, Urbana, IL 61801

chien@cs.uiuc.edu

Abstract

This paper analyzes the utility of using incremental reasoning to reduce the computational expense of constructing plan explanations for explanation-based learning. In particular this paper analyzes an approach in which an initial explanation is constructed considering a limited subset of all possible operator effects. This limited consideration corresponds to a set of non-monotonic persistence simplifications. Later incorrect predictions made by the explanations can then be used to direct consideration of previously unconsidered operator effects. This paper focusses upon comparing this incremental approach to plan explanation to the conventional approach of exhaustive reasoning about operator effects in three ways: 1) completeness and soundness properties; 2) computational complexity analysis; and 3) ongoing empirical evaluation.

Introduction

In real-world domains, large amounts of knowledge are needed to adequately describe world behavior. With the requisite complex domain theory, complete reasoning becomes a computationally intractable task. Even in game domains, such as chess, the combinatorics of brute-force computation are intractable [Tadepalli89]. Unfortunately many AI techniques such as planning and explanation-based learning [DeJong86, Mitchell86] involve construction of explanations, and hence reasoning. In Explanation-based Learning, this problem of a domain theory with a high computational cost is called the Intractable Domain Theory Problem [Mitchell86].

One method of dealing with this problem is to use simplified explanations. In our particular approach [Chien89a], these explanations are used to perform explanation-based learning of plans to learn a general partially-ordered plan to achieve a goal. In this approach, a system is given weak methods knowledge and heuristic simplifying assumptions. When constructing plan explanations to perform explanation-based learning of plans the system uses limited inference and these potentially unsound (non-monotonic) simplifications to reduce the complexity of the explanation process. This limited inference involves checking only a small subset of the possible subplan interactions. Because of this limited inference used in explanation learned plans are not necessarily correct. Since simplifications are caused by the limited inference, when a plan incorrectly predicts goal achievement it must be due to limited inference. When the system observes or experiences a plan failure it constructs an explanation for the failure and uses this explanation to refine the plan to avoid the failure in the future. This refinement represents the checking of an inference path missed in the initial analysis due to inference limitations.

By using feedback from plan execution as direction the system avoids the computationally intractable blind search for interactions inherent in explanation construction in a complex domain. Additionally, because the system has a known instance of the failure to explain, the process of determining faulty simplifications is facilitated.

While this approach has a strong intuitive appeal, relatively little work has been directed towards concretely justifying the benefits of the simplification-based approach. This paper focusses upon exactly that area, namely formalizing and quantifying the strengths and weaknesses of our particular simplification-based approach to reasoning. This analysis compares the simplification-based approach to that of conventional exhaustive reasoning in three ways:

1. completeness and soundness properties of the two approaches are discussed
2. computational complexity properties of the two approaches are described
3. an ongoing empirical evaluation of the two approaches is outlined.

The remainder of the paper consists of three main sections. In the first section, we describe our simplification-based approach to reasoning, including a short example. Next our simplification-based approach to reasoning is compared to exhaustive reasoning based upon the three criteria described above. Finally, we discuss related work and summarize the results of this paper.

Overview

Our incremental approach to explaining plans consists of four steps:

1. *Initial Learning:* The system learns an initial plan based upon a simplified explanation constructed using limited in-

ference. This initial simplified explanation can be learned from observation (as described in [Chien89a]) or can be constructed using a problem-solving component (as described in [Chien89b]).

2. *Expectation Violation:* Our approach uses expectation violations [Schank82] to indicate flawed plan explanations. There are two types of expectation violations. *Unexpected failures* can result from problem-solving or observation and occur when a plan explanation for goal achievement (either observed or constructed by the planner) is applicable but the goal is not achieved. An *unexpected success* occurs when the system observes a plan from its plan library and predicts failure (due to an applicable failure explanation attached to the plan in the plan library) but observes the plan to succeed.
3. *Explanation of Expectation Violation:* The system constructs an explanation of the violated expectation.
4. *Knowledge Modification:* The system analyzes the explanation of the violated expectation to determine which simplification from the initial explanation is incorrect and corrects the plan explanation via a process which removes the simplification.

This approach to constructing and refining explanations has been tested by implementation of a prototype refinement system. This refinement system uses a representation based upon situation calculus which allows representation of conditional effects of operators similar to [Pednault88]. This system constructs initial plan explanations considering only a reduced set of operator effects when checking for interactions between subportions of the plan. Because considering the complete set of operator effects is a computationally expensive task, in our approach, a system considers subplan interactions as directed by expectation violations. For a more detailed description of our initial learning and refinement approach see [Chien89a].

An Example

In order to clarify the plan refinement process, a simple example from a mission planning domain will be described. In this example, the system is given the goal of getting a certain amount of military force to a goal location (where military force present depends upon the number and type of units at the location and their readiness and supply state). The system constructs a plan which uses air transport for a number of airmobile infantry units to move these units to an intermediate airfield and then moving them by ground a short distance to the goal location. A number of support units are also moved entirely by ground to the final location. The system expects that this plan achieves the goal of getting the goal amount of force to the goal location.

This initial plan works for the current problem instance and correctly generalizes the plan to many other situations. For example, the exact airport used for the air transport can be generalized within distance constraints of the air transport, the exact units are not critical (although the unit types are), and the goal location is generalized (although it must be near the airfield and the starting locations of the support units).

However, in this case incomplete checking for interactions causes a faulty plan. The system incorrectly believes that this plan will work for any time of day of operation and only requires normal readiness and supply status for the airmobile infantry.

The system next attempts to use this plan in a case where the ground travel for the airmobile infantry unit takes place at night. The plan is executed and the infantry reaches the final location but at a decreased readiness and supply status, which results in an inadequate amount of force level at the goal location. The system queries the simulator to determine the causes for this failure. The system explains the failure as follows. The airmobile unit suffered a reduction in readiness from prolonged air travel. This low readiness was further reduced by a night maneuver (the ground movement from the airfield). The night maneuver additionally caused greater than expected supply expenditures because it took place at night. The factors together produced a readiness reduction in the infantry unit sufficient to cause the goal to fail.

This failed plan is then repaired by an analysis of how the causes of the failure could be prevented. First, the system notes that the failure depends upon the fact that the ground movement occurred at night. In cases where the movement can be scheduled during daylight hours the failure can be prevented. Second, the system notes that the failure is a reduction in the strength at the final location and notes that using a unit with a higher intrinsic strength will still allow the goal of having the desired strength at the final location to be achieved. Consequently the plan is modified to state that when a night ground movement is specified by the plan a higher strength unit is required.

Evaluation

This section contrasts the incremental planning approach described in this paper with the more conventional exhaustive planning approach. This comparison examines three aspects of each approach: 1) soundness/completeness guarantees of each approach, 2) computational complexity evaluations of computational cost, and 3) ongoing empirical evaluations of computational cost.

Soundness and Completeness

In general, planning involves constructing an explanation/proof that a set of actions will achieve a goal state. Because we assume a correct domain theory, soundness of a planning procedure means that any plans constructed using this procedure are guaranteed to work. In order to produce a sound explanation a system must search all of the potentially relevant rules and check them. For example, proving that a fact F persists from situation A to situation B through the execution of operator O1 would involve checking that all of the possible effects of O1. This would include performing all of the inferences to compute the new situation B. For example suppose F is (alive Fred) and O1 is (drive Fred home work). Exhaustively proving the persistence of (alive Fred) would require proving that numerous unlikely events do not occur. Due to the lack of common sense knowledge, an exhaustive reasoning system would have to prove that the car seats would not explode when warmed by the body heat of a passenger. And investigating just one of these possibilities is computationally quite expensive.

Consider investigating the exploding car seat possibility. Dismissing this possibility requires determining the maximum temperature of the seat caused by the body temperature and the passenger compartment temperature, the combustion temperature of the seat, and many other factors. In general because the exhaustive reasoning approach examines every possible reasoning path it must follow many potential reasoning paths that do not influence the final explanation. However, due to this exhaustive consideration of possible proofs an exhaustive reasoner can guarantee soundness of its plans.

Next consider the property of completeness which we define as guaranteeing that if a solution exists the planning procedure will find it. This means that a procedure attempts every method of goal achievement so it considers every spot in the search space. Exhaustive reasoning planners [Chapman87, Pednault88] can guarantee completeness in planning as we have defined it.

While conventional exhaustive reasoning planners guarantee completeness and correctness our incremental planning approach cannot guarantee soundness of an initial explanation because our procedure intentionally does not check all potential inferences. However, it can guarantee convergence upon soundness defined as follows:

Sound Model Convergence (soundness): As an initial plan is refined the predictiveness of the plan will eventually become exactly that of the exhaustive reasoning explanation and refinement will cease. Because refinements are triggered by incorrect plan predictions, the refinement process will occur as long as the incrementally learned plan makes incorrect predictions (i.e. predictions contradicting those made by the exhaustive reasoning plan). An initial plan is constructed using the same analysis process as the exhaustive reasoning approach except that it considers only a subset of the possible operator effects. With each refinement it adds for consideration the set of operator effects whose previous omission caused the incorrect predictions. When the set of operator effects considered becomes the set of actual relevant operator effects for the plan (i.e. those appearing in the exhaustive analysis) the predictions made by the plan will be the same as those made by the exhaustively derived plan. Because the plan contains a finite number of operators and each operator has a finite number of effects the total possible set of operator effects for the plan must be finite. Since the total possible set of operator effects is an upper bound on the set of actually relevant operator effects, the set of actually relevant operator effects must also be finite. Because each refinement adds a non-empty set of operator effects to the plan analysis and the number of operator effects needed for correct prediction is finite, eventually the refinements will lead to a plan predicting the same as the exhaustive reasoning plan.

Completeness: For a given class of input examples E, if there exists an explanation whose complete model correctly predicts goal achievement over E, the system will eventually generate such an explanation. This property relies upon the sound model convergence property described above. Our plan explanation method considers all of the methods of goal establishment (i.e. via direct operator effect, conditional operator effect, persistence from initial state). Our method also considers all of the operator effects for goal establishment. Hence any explanation for goal achievement can be generated in incomplete form as an initial explanation. Because there are a finite number of operators in the domain theory, a finite number of effects per operator, and a finite initial state, there are a finite number of plans to achieve a goal using a plan of a set number of operators. Let N be the number of operators in the shortest plan correctly predicting goal achievement over our set of examples E. Generate plans explaining goal achievement in increasing number of operators. If a plan cannot be refined to correctly account for an example it predicts identical to an exhaustive explanation which does not cover E. This will eventually happen for any incorrect choice of initial explanation as guaranteed by the sound model convergence property. Discard this plan and generate the next larger plan. Because there are a finite number of plans for goal achievement of size less than or equal to N we will eventually arrive at the shortest plan which can correctly predict for E and refine it to the sound predictiveness.

As a result of these properties, a system using our refinement approach is guaranteed to eventually produce a correct solution if one exists.

Computational Complexity Analysis

As shown above, our approach guarantees convergence upon soundness and completeness. However, because the main motivation for our approach is computational efficiency, a direct comparison of the computational expense of incremental reasoning and exhaustive reasoning is now discussed.

In general the cost of constructing a plan explanation consists of two parts: establishments – ensuring that facts that you want true are true at or before the time that you need them to be true; and protection – ensuring that facts stay true from the time they are established until the time they are needed. Because our approach performs exhaustive reasoning about establishments in order to retain completeness, the computational expense of reasoning about establishments is the same in the incremental and exhaustive reasoning approaches.

The cost of checking protections is the cost of checking each possible effect of an operator to see if it negatively affects any facts pertaining to the successful completion of the plan. These facts pertaining to the successful completion of the plan are called protected facts and include facts used as preconditions of operators (including conditional preconditions for conditional effects used in the plan) and facts used to satisfy the goal in the final state. Any effect that could potentially falsify a protected fact requires that the plan be constrained to prevent such a falsification. Because we use a partially-ordered plan representation, each of these checks is a expensive action. This is because determining the exact context in which the operator will be executed involves determining the truth value of facts in a partially ordered plan with conditional effects (an NP-hard problem [Chapman87]). Let:

E be the average number of effects per operator

C be the cost of determining whether an effect e can occur in a particular context of a given plan and if so under exactly

what conditions
P be the number of preconditions per operator
G be the number of items in the goal criterion
N be the number of operators in the plan

The number of protected facts in a plan is proportional to the number of preconditions of operators in the plan plus the number of facts in the goal specification or O(NP+G). The number of effects to check against these protected facts is O(NE). The total cost of checking protections in the exhaustive approach is the product of the protected facts and the effects times the cost of checking a single protection for a total expense of:

$$O(ENC(NP+G))$$

In the worst case C will be the cost of constructing a separate explanation for each possible ordering of each possible set of operators preceding the operator whose effect we are determining. This is at worst $O(2^N N!)$ explanations. The cost of constructing a support explanation for a single ordering would be $O(K^N)$ where K is the branching factor of the domain theory. In general the number of possible contexts to investigate would be much less than $O(2^N N! K^N)$ as this presumes totally unordered operators. Additionally, heuristics for preferring certain orderings [Drummond88] offer promise in reducing the number of orderings for consideration.

In contrast, the computational cost of our incremental approach depends upon the actual number of interactions occurring in the plan. This is because in our approach a protection is not checked until an incorrect plan prediction upon an example indicates that a protection violation can actually occur for a given effect–protection pair. Let:

X be the actual number of interactions occurring in the plan. Note that $X \leq EN(NP+G)$.

C' be the cost to determine the exact circumstances under which an effect e will occur in a plan given an example of the effect occurrence. Note that in most cases C' will be less that C because we can construct an explanation using the particular concrete failure example we have observed. If this explanation of the effect is of size S, the cost of constructing the explanation without guidance (as required by the exhaustive approach) is $O(k^S)$ where k is the branching factor in the domain theory. This is because there are S choices of k alternatives in the search space. In contrast, if there are T relatively uniformly spaced intermediate points in the concrete example to constrain the explanation process the cost of deriving the same explanation in the simplify and refine approach is $O(Tk^{S/T})$.

Thus the computational cost of the simplify and refine approach is: O(XC').

However, the simplify and refine approach also requires a number of examples to converge upon a correct plan. Because we assume that the system has the capability to isolate a faulty simplification in a single example the simplify and refine approach will fail on X examples(where again X is the actual number of interactions) before convergence upon a correct concept. If multiple isolatable failures occur in single examples the system can refine all of them and the system will require less than X failures.

To summarize the computational complexity results, the incremental reasoning approach provides significant computational savings over the exhaustive approach if: either 1) the actual number of effect–protection interactions is much less than the potential number of effect–protection (e.g. $X << EN(NP+G)$) or 2) the examples of actual interactions provide significant guidance through intermediate points in the explanation (e.g. C' is significantly less than C). However, these savings are contingent upon the ability to use failure instances to isolate faulty simplifications.

Empirical Evaluation (in progress)

This section describes ongoing empirical evaluation of the computational savings from using incremental reasoning in explanation construction. This evaluation consists of using a planner to solve mechanically generated problems. The first set of experiments involves using hand–coded domain theories operating in: 1) a simple workshop domain involving a drill, roller, and oven; and2) a mission planning domain involving simple logistics. The second set of experiments involves the use of machine–generated domain theories from the parameters of E = # of effects per operator and P/E = # of preconditions per operator effect. These experiments will provide empirical figures on:

The # of potential interactions in a plan and the # of actual interactions and how this figure is affected by E and P/E. We have shown analytically that increasing P or E increases both the potential and actual # of interactions. Additionally, increasing P/E should decrease the percentage of potential interactions which turn out to be actual interactions.

the # of nodes searched in verifying an interaction via the exhaustive approach & the # of nodes searched in verifying an interaction via the incremental approach. This will empirically measure the values of C and C'.

the # of examples required by the incremental approach to converge upon a sound concept and the relation of this to the error rate of the concept.

One goal of these empirical tests is to attempt to derive a static test which will indicate the expected performance of an our incremental reasoning approach to an arbitrary domain theory. This static test would use the E and P/E properties of the domain theory to predict the expected costs of applying the incremental and exhaustive reasoning approaches.

Because these tests are not yet complete, we can offer only tentative empirical results. These results indicate that increasing P/E strongly influences the ratio of potential to actual interactions (the ratio $X/(EN(NP+G))$. These initial empirical results also indicate that the number of intermediate guidance points gained by the incremental reasoning approach in investigating an actual interaction is O(N). This information is of interest because it determines the ratio of C' to C.

Discussion and Summary

While there have been a number of problem–solvers based upon the approximate and refine approach to problem–solving [Gupta87, Hammond86], relatively little work has addressed directly comparing the approximate and refine approach to the exhaustive approach. Unruh and Rosenbloom [Unruh89] present interesting empirical results but do not provide a computational complexity analysis. Tadepalli presents an analysis of his partial reasoning approach to constructing explanation for EBL [Tadepalli89] and compares it to an Alpha–Beta approach. However, Tadepalli's approach is designed for a two–player adversarial game situation whereas our work concerns single–agent planning. Simmons [Simmons88] also uses an approximate reasoning approach to problem solving which he calls Generate–test–debug. However, the complexity of certain types of reasoning within Simmons' approach makes a direct analysis of GTD versus an exhaustive problem–solver difficult. However, Simmons does provide a characterization of which domain he feels his approach will be applicable ([Simmons88] Section 6.3). Despite this relatively sparse incidence of detailed analyses we feel that the type of analysis we have performed on our approach to incremental reasoning can be performed, in principle, upon the other incremental reasoning approaches mentioned.

One interpretation on this work is that it addresses *The Frame Problem* [McCarthy69] in that it deals with efficiently reasoning about the myriad of possible changes caused by the execution of operators in a complex environment. Our approach to incremental reasoning indicates that by using experience to guide the investigation of relevant effects the computationally intractable blind search of potentially relevant effects can be avoided.

Conventional exhaustive planners for conditional effects are another area of related work. However, these systems must use carefully crafted restricted representation languages or face serious computational complexity problems. For example, Pednault's approach [Pednault88] requires that all of the conditions under which a fact is known to persist through the execution of an operator be explicitly stated (preservation conditions). Consider attempting to state all of the conditions under which driving your car does not end the fact that the driver is alive (and the difficulty of reasoning about the large number of disjuncts). These representational restrictions are a major factor reducing the usefulness of these systems.

This paper has analyzed the computational gains of an incremental reasoning approach applied to constructing plan explanation. In this incremental reasoning approach the system performs limited considerations of operator effects when constructing a plan explanation. The system uses later incorrect predictions by the plan to direct expansion of the initial plan into a correct–predicting plan. Our analysis showed that our approach guarantees convergence upon a sound plan and complete coverage of the solution space. A computational complexity analysis then showed the computational gains of the incremental reasoning approach to be in cases where: 1) the number of actual interactions is small but the number of potential interactions is large; and 2) the direction given by the incorrect prediction provides significant guidance in determining the cases in which a previously unconsidered effect is relevant. Finally, we discussed an ongoing empirical evaluation of our approach to incremental reasoning which is directed at providing strong information upon the applicability of our approach.

Acknowledgements. Comments and direction from Gerald Dejong are gratefully acknowledged. This research was supported by an IBM Graduate Fellowship and The National Science Foundation under grant NSF–IRI–87–19766.

References

[Chapman87] D. Chapman, "Planning for Conjunctive Goals," *Artificial Intelligence 32*, 3 (1987), pp. 333–378.

[Chien89a] S. A. Chien, "Using and Refining Simplifications: Explanation–based Learning of Plans in Intractable Domains," *Proceedings of The Eleventh International Joint Conference on Artificial Intelligence*, Detroit, MI, August 1989, pp. 590–595.

[Chien89b] S. A. Chien, "Failure–guided Search in Planning," *Proceedings of the American Association for Artificial Intelligence Spring Symposium on Planning and Search*, Palo Alto, CA, March 1989.

[DeJong86] G. F. DeJong and R. J. Mooney, "Explanation–Based Learning: An Alternative View," *Machine Learning 1*, 2 (April 1986), pp. 145–176.

[Drummond88] M. Drummond and K. Currie, "Exploiting temporal coherence in nonlinear plan construction," *Computational Intelligence 4*, (1988), pp. 341–348.

[Gupta87] A. Gupta, "Explanation–based Failure Recovery," *Proceedings of the National Conference on Artificial Intelligence*, Seattle, WA, July 1987, pp. 606–610.

[Hammond86] K. Hammond, "Learning to Anticipate and Avoid Planning Failures through the Explanation of Failures," *Proceedings of the National Conference on Artificial Intelligence*, Philadelphia, PA, August 1986, pp. 556–560.

[McCarthy69] J. McCarthy and P. J. Hayes, "Some Philosophical Problems from the Standpoint of Artificial Intelligence," in *Machine Intelligence 4*, B. Meltzer and D. Michie (ed.), Edinburgh University Press, Edinburgh, Scotland, 1969.

[Mitchell86] T. M. Mitchell, R. Keller and S. Kedar–Cabelli, "Explanation–Based Generalization: A Unifying View," *Machine Learning 1*, 1 (January 1986), pp. 47–80.

[Pednault88] E. Pednault, "Synthesizing plans that contain actions with context–dependent effects," *Computational Intelligence 4*, (1988), pp. 356–372.

[Schank82] R. C. Schank, *Dynamic Memory*, Cambridge University Press, Cambridge, England, 1982.

[Simmons88] R. Simmons, "Combining associational and causal reasoning to solve interpretation and planning problems," Technical Report 1048, Artificial Intelligence Lab, Massachusetts Institute of Technology, Cambridge, MA, 1988.

[Tadepalli89] P. Tadepalli, "Lazy Explanation–Based Learning: A Solution to the Intractable Theory Problem," *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, Detroit, MI, August 1989.

[Unruh89] A. Unruh and P. Rosenbloom, "Abstraction in Problem Solving and Learning," *The Eleventh International Joint Conference on Artificial Intelligence*, Detroit, MI, 1989, pp. 681–687.

A GENERAL THEORY OF ABDUCTION

Kurt Konolige
Artificial Intelligence Center *and*
Center for the Study of Language and Information
SRI International
Menlo Park, CA

Introduction

We present a general theory of abduction. This theory is intended to formalize a notion of abduction within a logical framework that is general enough to represent the typical uses of abduction in Artificial Intelligence, e.g., diagnosis, explanation, plan recognition, and so forth. The main features of the theory are:

1. The generalization of the logic of abduction to incorporate default assumptions.
2. A clear separation of the role of default assumptions from the abductive assumptions necessary to explain observations.
3. Analysis of the relation of the theory to nonmonotonic formalisms that have been used for abduction; in particular we show that a subset of the theory can be treated as a closure and minimization operation, using default logic or circumscription. This result shows the relation between consistency and abduction-based treatments of diagnostic inference.
4. Implementation of the theory by a Doyle-style (justification-based) TMS. The implementation is exact only for a restricted form of the theory. We show how the TMS generalizes the ATMS in an abductive framework.

Prediction and Explanation

The general framework we assume is that there are causal relations among events in the world, and we can model the world by representing and reasoning about these relations. There are two basic types of reasoning operations:

1. Prediction of effects from causes, and
2. Assumption of causes from effects.

For example, in reasoning about action, we may know that the sprinklers were turned on last night, and so predict that the lawn will be wet in the morning. This is the so-called *temporal projection* problem: how to infer the consequences of a set of actions in an initial situation. On the other hand, we may observe that the lawn is wet, and so infer that the sprinklers were turned on. This assumption explains the observation by giving a cause for it.

The prediction of effects from causes often demands the use of defaults, since the knowledge of a situation may be imperfect. The lawn may not be wet even after the sprinklers are turned on, because they may fail to work properly. Through the use of defaults, it is possible to state the conditions that normally would be caused by an action: the best we can do in complex domains. Such defaults are obviously defeasible, because if better information becomes available, the initial default conclusions may be retracted.

The assumption of causes from effects is abductive in nature. Roughly speaking, we seek the best explanation for the observed effects. Abductive reasoning is obviously defeasible: knowing the lawn is wet might be sufficient evidence to conclude that the sprinkler was on as the best explanation (especially in a dry climate). Further knowledge that it had rained would make this conclusion unfounded.

The fact that both prediction of effects and assumption of causes are defeasible can lead to a confusion of the two in a formal account of reasoning. The approach we take in the next section distinguishes them clearly.

A Logical Theory of Abduction

The account of abduction that we are interested in has been termed "logic-based". That is, the causal relations among events in the world are treated as a theory in some logical framework, and observations and as-

sumptions are expressed as sentences in the logic This approach is weaker than probabilistic accounts in its ability to order the plausibility of assumptions, it is stronger in its ability to represent complex domains.

Within the logical approach, there have been many different accounts of abduction, some with respect to particular domains (e g., [Reiter, 1987] for diagnosis), others of a more general nature (e.g., [Poole, 1988]). The account we give here draws on ideas from these, and formalizes them in a general way. The abductive inference problem is stated with respect to an *abductive frame* giving the appropriate language and background theory.

DEFINITION 1 *An* abductive frame *is a tuple* $\langle \mathcal{L}, \Sigma, \mathcal{A}, \mathcal{O} \rangle$, *where*

- $\mathcal{L}$ *is a logical system.*
- Σ *is a set of sentences of* $\mathcal{L}$, *the* background theory.
- $\mathcal{A}$ *is the assumption vocabulary.*
- $\mathcal{O}$ *is the observation vocabulary.*

The logical system is arbitrary, as long as it has a well-defined notion of consequence, which we express by $\Gamma \vdash_{\mathcal{L}} \phi$, that is, the sentence ϕ follows from the set of sentences Γ in the system $\mathcal{L}$. The background theory expresses knowledge of the causal relations of the world. The observation and assumption vocabularies are used to express what we observe about the world and what we are willing to assume to explain these observations.

For example, consider the domain of reasoning about action. We'll choose the logical system to be default logic. The background theory expresses knowledge of the way in which actions change the world. For the lawn example, we might have:

$$\forall t\ \mathit{rain}(t) \supset \mathit{wet\text{-}lawn}(\mathit{suc}(t))$$
$$\forall t\ \mathit{rain}(t) \supset \mathit{wet\text{-}road}(\mathit{suc}(t))$$
$$\forall t\ \mathit{sprinkler}(t) \supset \mathit{wet\text{-}lawn}(\mathit{suc}(t))$$
$$\forall t\ \mathit{sun}(t) \supset \mathit{dry\text{-}road}(\mathit{suc}(t))$$

$$\mathit{wet\text{-}lawn}(t) : \mathit{wet\text{-}lawn}(\mathit{suc}(t)) / \mathit{wet\text{-}lawn}(\mathit{suc}(t))$$
$$\mathit{dry\text{-}lawn}(t) : \mathit{dry\text{-}lawn}(\mathit{suc}(t)) / \mathit{dry\text{-}lawn}(\mathit{suc}(t))$$
$$\mathit{wet\text{-}road}(t) : \mathit{wet\text{-}road}(\mathit{suc}(t)) / \mathit{wet\text{-}road}(\mathit{suc}(t))$$
$$\mathit{dry\text{-}road}(t) : \mathit{dry\text{-}road}(\mathit{suc}(t)) / \mathit{dry\text{-}road}(\mathit{suc}(t))$$

$$\forall t\ \mathit{dry\text{-}lawn}(t) \equiv \neg \mathit{wet\text{-}lawn}(t)$$
$$\forall t\ \mathit{dry\text{-}road}(t) \equiv \neg \mathit{wet\text{-}road}(t)$$

This is a simple situation-calculus theory, with properties like *wet-lawn* being true in a given situation, and events like *rain* occuring at a given situation and having effects in the succeeding one. The persistence of properties from one situation to the next is represented by the default rules. Given an initial situation in which the lawn is dry and no events occur, for example, we would conclude that the lawn is still dry in succeeding situations.

In a simple form of temporal projection, the sequence of events and (perhaps) some information about the initial situation is given. The assumption vocabulary consists of properties of the initial situation, since these are the causes (along with the events) of properties in subsequent situations. The observational vocabulary contains the properties of situations after the initial one, since these are to be explained.

In a typical case, we might know:

$\mathit{suc}(0) = 1, \mathit{suc}(1) = 2$	given
$\mathit{rain}(1)$	given
$\mathit{dry\text{-}road}(0)$	given
$\mathit{dry\text{-}lawn}(1)$	observation

From this initial information, we can predict from the background theory alone that the road will be dry in situation 1 and wet in situation 2. To explain the observation of the lawn being dry in situation 1, we will have to assume that it was dry in situation 0.

We now state the general form of abductive reasoning.

DEFINITION 2 *Let* $\langle \mathcal{L}, \Sigma, \mathcal{A}, \mathcal{O} \rangle$ *be an abductive frame. Let* S, *the* situational facts, *be a finite set of sentences not containing the vocabulary of* $\mathcal{O}$. *Let* O, *the* observations, *be a sentence from the vocabulary of* $\mathcal{O}$.

An explanation *of the observations is a finite set* $A \subset \mathcal{A}$ *such that*

1. $\Sigma \cup S \cup A \vdash_{\mathcal{L}} O$.
2. A *is consistent with* Σ.
3. A *is minimal.*

A cautious explanation *is the disjunction of all the explanations, that is,* $\bigvee_i A_i$.

For the example above we have

$$S = \{\mathit{rain}(1), \mathit{dry\text{-}road}(0), \mathit{suc}(0) = 1, \mathit{suc}(1) = 2\}$$
$$O = \mathit{dry\text{-}lawn}(1)$$

The only possible explanation is $\mathit{dry\text{-}lawn}(0)$, which thus must be the cautious explanation. From the background theory the conclusions are $\mathit{dry\text{-}lawn}(1)$, $\mathit{dry\text{-}road}(1)$, $\mathit{wet\text{-}road}(2)$, and $\mathit{wet\text{-}lawn}(2)$, and corresponding negative literals.

Remarks. The situational facts are meant to include any knowledge of the particular situation at hand that does not need to be explained. We could include such information as part of the background theory, but it is more convenient to separate it, since an abductive frame will be useful over many particular situations.

The restrictions on the form of A deserve comment. A must be a minimal set of expressions from the assumption vocabulary $\mathcal{A}$; by minimal is meant there is no other explanation that is a proper subset The idea is that these expressions constitute valid causes of the observed effects, and if there is an explanation which contains fewer causes, it should be preferred. Other than this we say nothing about preferences among multiple explanations. It is obvious that often such preferences will be required for reasoning, e.g . we may want the most specific explanation, or the most probable, or the X-est, where X is some measure on explanations. The preference could be expressed mathematically by a partial order on the subsets of $\mathcal{A}$. Since such an order will be closely related to the domain of application, and we have no way of making any general statements about the order, we omit it from further consideration here.

In a given problem domain, we may be interested in the best explanation, or the cautious explanation, or even any (satisficing) explanation. For example, if we want to predict the possible states of the world under a series of events, then the cautious explanation might be most appropriate. Tasks like plan recognition usually require the best explanation. And for some problems, like the N-queens problem, there is no ordering of solutions, and any one would be acceptable.

This theory of abduction presented here is a general one because the logical system $\mathcal{L}$ and the background theory are not restricted to a particular type of first-order theory, as is often done. In particular, we are free to use a nonmonotonic logic for $\mathcal{L}$, to express default predictive conclusions in a causal theory, as in the example above.

Implementation using a TMS

The definition of abduction just given is very general, and if it is to be used as the basis for building reasoning systems, there must be a proof theory and implementation. In the general case, the problem of deriving explanations is (a) r.e. when $\mathcal{L}$ is first-order, and (b) non-r.e. when $\mathcal{L}$ is default logic. For finite propositional languages, both these cases are decidable, although the complexity may be high. What we will do here is relate the abductive theory (in a restricted propositional case) to the ATMS [de Kleer, 1986] and TMS [Doyle, 1979], and then show how this relationship can be generalized to yield an approximate procedure for the general case.

From the results of [Reiter and de Kleer, 1987], it is easily shown that the ATMS can compute all explanations of observations O for a background theory and situational facts $\Sigma \cup S$, when O is a single (positive) atom and Σ and S can be expressed as Horn clauses. In fact, even without the restriction to Horn clauses, the ATMS could be used to compute explanations, as long as all the Horn clauses relevant to the observations were derived from the background theory and added to the ATMS.

The ATMS is not sufficient when $\mathcal{L}$ is default logic.[1] Recent work with Ulrich Junker [Junker and Konolige, 1990] has shown that the TMS can be regarded as an implementation of default logic. In particular, when the first-order part of a default theory is Horn, there is a simple mapping to a TMS such that the extensions of the default theory correspond to the extensions (admissible labelings) of the TMS. We will use this results, and show how to generalize the notion of extension of a TMS to generate explanations.

We take the TMS to be the formal version given in [Reinfrank and Freitag, 1988], but extend it by adding a special symbol $\perp$ for contradiction. Nodes of the TMS are atoms (including $\perp$). A justification is of the form $(M|N \rightarrow c)$, where M is a set of nodes (the monotonic antecedents), N is a set of nodes (the nonmonotonic antecedents), and c is a node (the conclusion). A TMS theory is a set of justifications.

Informally, a node n is provable from a set of nodes E in a theory J if there exists a noncircular application of justifications of J, leading to n, that are valid in E. E is grounded if every node in it is provable from it. It is closed if it contains the conclusion of every valid justification of J. An extension of J is any set that is both grounded and closed.

From the results of [Junker and Konolige, 1990], we can show the following connection between extensions of a default theory and a corresponding TMS theory.

DEFINITION 3 *Let W be a set of Horn clauses of the form $p_1 \wedge \cdots \wedge p_n \supset q$, where q may be $\perp$. Let D be a set of defaults $a_1 \wedge \cdots \wedge a_m : b/c$ where a_i, b, and c are atoms. The default theory $\langle W, D\rangle$ is called a Horn default theory. The corresponding TMS theory is given by the set of justifications $\langle\{p_1, \cdots, p_n\}|\emptyset \rightarrow q\rangle$ and $\langle\{a_1, \cdots, a_m\}|b \rightarrow c\rangle$ formed from W and D.*

THEOREM 1 *Let $\langle W, D\rangle$ be a Horn default theory, and J the corresponding TMS theory. Let $\mathrm{Cn}(E)$ be the propositional consequences of the set E. Then E is an extension of J if and only if $\mathrm{Cn}(E)$ is an extension of the default theory $\langle W, D\rangle$.*

[1] There have been recent attempts to generalize the ATMS algorithm to the nonmonotonic case [Reinfrank *et al.*, 1989, Junker, 1989], but there are still problems; in particular, the language is restricted so that contradiction is excluded. Here we view the TMS as a generalization of the ATMS, and allow contradiction (or NOGOODS in the terminology of [de Kleer, 19[illegible]

The TMS, while able to compute default extensions for Horn default theories, differs from the ATMS in that it does not compute the minimal assumption sets for which a given conclusion would hold in the theory. This is what we require. We generalize the definition of extension along the lines of [Reinfrank *et al.*, 1989], by introducing a set $A \subset \mathcal{A}$ of assumptions that are unjustified. E is an A-extension of J if $E \cup A$ is closed and every node of E has a proof in $E \cup A$. Note that this differs from the previous definition of extension only in that the nodes of A do not need proofs. An *explanation* of a node n in theory J is a minimal assumption set A such that n is a member of some A-extension of J. The following theorem shows the essential equivalence of abductive theory explanations and TMS explanations.

THEOREM 2 *Let $\langle \mathcal{L}, \Sigma, \mathcal{A}, \mathcal{O} \rangle$ be an abductive frame, and let $\Sigma \cup S$ be a Horn default theory with J its corresponding TMS theory. Then A is an explanation for $o \in \mathcal{O}$ if and only if A is an explanation for o in J.*

To pursue the example given in the last section: although it is not in the form of a Horn default theory, minor modifications will make it so. First, replace all the universally-quantified statements with their instantiations for situations 0, 1, and 2, e.g.,

$$\mathit{rain}(0) \supset \mathit{wet\text{-}lawn}(1) .$$

Next, change the equivalences to simple contradiction, e.g.,

$$\mathit{dry\text{-}lawn}(0) \wedge \mathit{wet\text{-}lawn}(0) \supset \bot .$$

Now the theory is Horn, and can be translated into a justification network. The only explanation containing $\mathit{dry\text{-}lawn}(1)$ has $A = \{\mathit{dry\text{-}lawn}(0)\}$, just as for the abductive theory.

At this point, we have shown that a suitably defined TMS is a generalization of the ATMS, allowing nonmonotonic justifications. We are left with two tasks: to extend the TMS translation to non-Horn default theories, and to construct algorithms for computing explanations of TMS theories. In [Junker and Konolige, 1990], the first results on extending the translation to less restricted default theories are given; in general the translated TMS will be only an approximation of the default theory.

For the second task, we can take advantage of a large body of research on constraint-satisfaction methods. We have been using a type of forward-checking algorithm. Although in a preliminary stage of analysis, this algorithm has proven to be efficient in a variety of tasks, especially where only a satisficing explanation is necessary. We have concentrated mostly on problems in which the constraints are difficult to satisfy, but where the abduction is straightforward. A typical example is the N-queens problem Here the algorithm produces a solution to the 10-queens problem every 0.1 seconds, and a solution to the 500-queens problem every 12 seconds.

Closure + Minimization implies Abduction

In this section we prove a result about the relation of general abduction to closure and minimization for causal theories. There are two distinct logic-based approaches to diagnosis, a type of abductive task using a causal theory. In one, the abductive approach, a system is diagnosed by finding causes for the observed symptoms. This type of diagnosis is readily represented in the general abduction formalism presented here: the background theory contains the relation between causes and symptoms, and explanations give the diagnosis of observed symptoms. On the other hand, accounts of diagnostic reasoning have also been given in terms of minimization and logical consistency. For example, in [Reiter, 1987], a diagnosis is a minimal set of abnormalities that is consistent with the observed behavior of a system.

At first glance, these two approaches seem fundamentally different in at least two respects. First, as we noted already, the form of inference is distinct, since abduction from the background theory Σ to observations O is not the same as the consistency of Σ and O. Second, they encode knowledge of the domain in different ways: in the abductive framework, there are implications from the causes to the effects, while in the consistency-based systems, the most important information seems to be the implication from observations to possible causes. For example, in Reiter's reconstruction of the set-covering model of diagnosis [Reiter, 1987], he uses axioms of the form:

$$\textsc{observed}(m) \supset \textsc{present}(d_1) \vee \cdots \vee \textsc{present}(d_n),$$

where m is the observed symptom and d_i are diseases that cause the symptom.

While these differences exist, from a more abstract point of view there is a clear connection between the two approaches: they are different implementations of causal abduction. To see this, consider a particular element o of the observation vocabulary. In terms of the general abduction theory, o will have a set of explanations $A_1, A_2, \cdots, A_n$ relative to a background theory Σ. Let us call these the *causes* of o. Now suppose we add to the background theory a statement

$$o \supset A_1 \vee A_2 \vee \cdots \vee A_n ,$$

where we understand each A_i to be the conjunction of its elements. This expression says that whenever o is present, it must have been caused by one of the A_i. Let Σ' be the result of adding to Σ one such statement for each o. As a background theory, Σ' is much stronger than Σ, since it contains the closure over all possible causes for each observation.

Now suppose we observe o. This observation is consistent with Σ', and $A_1 \vee A_2 \vee \cdots \vee A_n$ is true in all consistent models of o and Σ'. If we now try to minimize causes, that is, to assert $\neg A_i$ (again understanding A_i as a conjunction) for as many causes as possible, we will eliminate possible causes from the disjunction, until we are left with a single cause. Thus we can perform abductive reasoning in the consistency-based approach, given closure over causes and minimization of causes. We make this more precise with the following theorem.

THEOREM 3 *Let $(\mathcal{L}, \Sigma, \mathcal{A}, \mathcal{O})$ be an abductive frame, with $\mathcal{L}$ first-order. Construct $\Sigma' \supset \Sigma$ as described above, by adding the closure of all causes for each observation. If A is an explanation for the observation sentence O, then for some maximal subset $X = \{\neg a | a \in \mathcal{A}\}$ such that $X \cup \Sigma$ is consistent with O, A is a logical consequence of $\Sigma' \cup X$ and O.*

The converse of this theorem is not true in general, since closure and minimization is a more powerful technique than abduction. For example, if a particular effect o is not observed and is not a consequence of the background theory, then the consistency-based system concludes its negation, while abduction does not.

The relation of general abduction to consistency-based approaches should now be fairly clear. In the latter case, for example with diagnostic systems such as Reiter's, or Kautz's theory of plan recognition [Kautz, 1987], abductive inferences are obtained by adding closure axioms to the background theory, and minimizing causes. From among the resulting explanations, still further refinements are possible: in Kautz's system causes which are minimal in cardinality are preferred; this corresponds to choosing a preference criterion on explanations in general abduction.

Whether it is preferable to use one or the other approach depends on the nature of the domain and the task. In those cases where reliable closure knowledge is not available, the consistency-based approach will force conclusions that are incorrect. Where this knowledge is available, it can lead to stronger conclusions than the abductive approach. Finally, the general abduction framework presented here has the advantage of integrating default reasoning about causation; to do the same for the consistency-based approach it would be necessary to introduce an ordering on defaults and the minimization of causes.

Acknowledgements. The research reported in this paper was partially supported by the NTT Corporation, and by the Office of Naval Research under Contract No. N00014-89-C-0095.

References

[de Kleer, 1986] Johan de Kleer. An assumption-based truth maintenance system. *Artificial Intelligence*, 28.127–162, 1986.

[Doyle, 1979] John Doyle. A truth maintenance system. *Artificial Intelligence*, 12(3), 1979.

[Junker and Konolige, 1990] Ulrich Junker and Kurt Konolige. Computing the extensions of autoepistemic and default logics with a TMS. In *Conference on Theoretical Aspects of Reasoning about Knowledge*, Asilomar, CA, 1990.

[Junker, 1989] Ulrich Junker. A correct non-monotonic ATMS. In *Proceedings of the International Joint Conference on Artificial Intelligence*, Detroit, Michigan, 1989.

[Kautz, 1987] Henry Kautz. A formal theory for plan recognition. Technical Report TR–215, University of Rochester, 1987.

[Poole, 1988] David Poole. A methodology for using a default and abductive reasoning system. Technical report, Department of Computer Science, University of Waterloo, Waterloo, Ontario, 1988.

[Reinfrank and Freitag, 1988] Michael Reinfrank and Hartmut Freitag. Rules and justifications, a uniform approach to reason maintenance and non-monotonic inference. In *Proceedings FGCS-88*. Tokyo, Japan, 1988.

[Reinfrank *et al.*, 1989] Michael Reinfrank, Oskar Dressler, and Gerhard Brewka. On the relation between truth maintenance and autoepistemic logic. In *Proceedings of the International Joint Conference on Artificial Intelligence*. Detroit, Michigan, 1989.

[Reiter and de Kleer, 1987] Raymond Reiter and Johan de Kleer. Foundations of assumption-based truth maintenance systems: preliminary report. In *Proceedings of the American Association of Artificial Intelligence*. Seattle, WA, 1987.

[Reiter, 1987] Raymond Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32, 1987.

A Theory of Abduction Based on Model Preference

Douglas E. Appelt
Artificial Intelligence Center
SRI International
Menlo Park, California

1 Introduction

A number of different frameworks for abductive reasoning have been recently advanced. These frameworks appear on the surface to be quite different. These different approaches depend on, for example, statistical Bayesian methods (see Pearl [4] for a survey), minimization of abnormality (Reiter [6]), default-based methods (Poole [5]), or assumption-based methods, in which unproved literals may be added to the theory as assumptions during the course of a proof (Stickel [9], Hobbs et al. [2]).

Although these abduction methods are grounded in the particular theories on which they are based, e.g., probability or default logic, there has not yet been a completely satisfactory theory of abduction in general that can account for the variety of reasoning and representation schemes encountered in all of these methods. The best effort to date in this direction has been undertaken by Levesque [3], who characterizes an abduction problem as finding all sets of explanations α for an observation β within a theory T. A proposition α is an explanation for β if $T \models (\alpha \supset \beta)$ and $T \not\models \neg\alpha$. Levesque alters this definition slightly by the introduction of a belief operator to T, which allows him to abstract from the particular rules of inference that may be used to conclude ϕ. He considers two possible definitions of the belief operator, each with different algorithms for computing assumptions that have different computational properties.

Within any abductive reasoning method there will generally be a set of assumptions, which could be used together with the theory to derive the desired conclusions. Levesque convincingly demonstrates that no purely semantic criterion can be used to distinguish competing assumptions, and proposes a syntactic metric based on the number of literals comprising the syntactic representation of the assumptions. This criterion will admit a number of competing explanations, each of which is minimal according to this criterion. Certainly in a large number of practical problems, one is very much interested in distinguishing a "best" explanation among all those that meet the syntactic minimality criterion. Typically such preferences depend on particular facts about the domain in question. It would therefore be desireable if there was some way of expressing domain-specific preference information within the theory so that syntactically minimal alternatives could be compared.

A number of proposals have been advanced for semantic criteria for comparing different sets of assumptions. For example, if the theory of a domain can be expressed naturally in terms of the normality and abnormality of the individuals in that domain, as is often the case with diagnostic problems, an obvious criterion to distinguish assumption alternatives is the number of abnormal individuals that are implied by the assumptions. Minimization of abnormality is a very natural preference criterion in such domains. However, not all abduction problems are best viewed in terms of abnormality of individuals. In fact, in natural-language processing, minimization strategies are quite inappropriate. If a speaker says, "My watch is broken," minimization strategies would consider why a typical speaker's own beliefs might support such an utterance. For exam-

ple, he might believe that the mainspring was broken, or perhaps a dozen different equally likely mental states. However, the hearer of such an utterance is really trying to infer what the speaker *intends him to believe*. In this case the intention is most likely reflected by the content of the utterance itself, i.e., the speaker's watch is broken, and not by any more specific cause that would support such a belief for the speaker. Stickel [9] proposes a different comparison criterion, which he calls *least specific abduction*, which is argued to be more appropriate for natural-language interpretation problems.

An alternative to abnormality-based approaches is to encode information about the desirability of different assumptions in the theory itself. In a Bayesian framework, this is expressed by the prior probabilities of the causes, and the probabilities of observations given causes. Another alternative, proposed by Hobbs et al. [2] involves encoding preferences among assumptions as weighting factors on antecedent literals of rules.

In this paper, I propose a model-theoretic account of abduction that represents domain-specific preferences among assumptions as preferences among the models of the theory. This proposal is directed toward the goal of developing a theory of abduction which characterizes domain-specific preference information abstractly, and which hopefully can be unified at some point with model theoretic accounts such as Levesque's. It is work in progress, and at this point consists more of definitions than theorems, but I believe the proposal is worthy of consideration in the search for a unified theoretical approach to abduction. I shall use the weighted abduction theory of Hobbs et al. [2] as an example of a possible computational mechanism to realize this approach.

2 A Theory of Abduction Based on Model Preference

Shoham [8] introduced the idea of model preference as a general way of expressing various forms of nonmonotonic inference. He postulates a partial preference order on the underlying models of a theory, and the desired conclusions of the theory are those propositions that are satisfied in all the maximally preferred models of the theory. In contrast with this global notion of preferential entailment, Selman and Kautz [7] introduce a logic they call *model preference default logic*, in which the individual default rules of the theory are interpreted as *local* statements of model preferences. For example, the default rule $p \rightarrow q$ is interpreted model-theoretically as a preference for models that satisfy q among all models that satisfy p.

If abductive reasoning is to be done within a theory, it is possible to give an interpretation to implications within that theory as expressing local preferences among models in a manner similar to Selman and Kautz's default rules. For example, if $p \supset q$ is a rule, and q is an observation, then the fact that p can be assumed as an explanation for q suggests an obvious model-preference interpretation of the rule: Among models satisfying q, models that satisfy p are "by and large" preferred to models satisfying $\neg p$.

The reason the hedge "by and large" is used in the above definition is that it cannot be the case that the abductive interpretation of $p \supset q$ is that, for all models that satisfy q, *every* model that satisfies p is preferred to *every* model that satisfies $\neg p$. It may be the case that other rules in the theory imply preferences that may be consistent with q, but inconsistent with p. In general, this criterion is too restrictive to permit the existence of a consistent model preference ordering for many theories of practical interest. A weaker interpretation of the relation between a rule and the model preference order is that every model satisfying p is prefered to *some* model satisfying $\neg p \wedge q$. Adding an assumption to a theory restricts the models of the theory. If this restriction is such that it rules out some models that are known to be inferior to every model of the theory plus the assumptions, and the theory plus the assumptions entails the observations, then the assumptions are a potential solution to the abduction problem. A set of assumptions A_1 is preferred to a set of assumptions A_2 for a given theory T, if every model of $T \cup A_1$ is preferred to some model of $T \cup A_2$. Abduction can thus be regarded as a problem of finding a set of assumptions that imply a greatest lower bound on the model-preference relation among other competing sets of assumptions.

A further possibility that needs to be considered is that, once an assumption set is found, there may exist models satisfying sets of assumptions that are inconsis-

tent with the assumption set under consideration, and every one of their models are preferred. Interpreted in terms of domain specific preferences, this would be a situation in which p is a possible explanation for q, but p and r cannot be true simultaneously, and r is almost always true. In such a situation, we say that the assumption of p is *defeated*, unless r can be ruled out by further preferred assumptions.

The following is a precise definition of abduction in terms of model preference.

Given a theory T, a total, antireflexive, antisymmetric preference relation $\succ$ on models of T, and an observation φ, an abduction problem consists in deriving a set of assumptions A that satisfies the following conditions:

1. **Adequacy.** $T \cup A \models \phi$
2. **Consistency.** $T \cup A \not\models \neg\phi$
3. **Syntactic minimality.** If $\psi \in A$ then $T \cup A - \{\psi\} \not\models \phi$
4. **Semantic greatest lower bound.** There is no assumption set A' such that:
   (a) $T \cup A'$ is adequate, consistent, and syntactically minimal
   (b) There exists $M \models T \cup A$ such that for every $M' \models T \cup A'$, $M' \succ M$
5. **Defeat condition.** There is no set A'' such that
   (a) There is some $\psi \in A$ such that $T \cup A'' \models \neg\psi$ and there is some $M \models T \cup A$ such that for every model $M'' \models T \cup A'', M'' \succ M$.
   (b) **Defeat exception.** There is no set of assumptions A''' such that
      i. if $M \models T \cup A'''$, then $M \models T \cup A$, and
      ii. there exists $M'' \models T \cup A''$ such that for every $M''' \models T \cup A''', M''' \succ M''$.

The adequacy and consistency requirements of this definition should be obvious. Because it may be possible to restrict the models of a theory to a favored subset by making assumptions that have nothing to do with the observation, the syntactic minimality problem imposes the requirement on the assumption set that every assumption must actually contribute to the solution of the problem. The greatest lower bound condition guarantees that the assumption set that constitutes the solution to the problem is one that is preferred to other assumption sets, provided that it is not defeated. An assumption set that is potentially defeated is still admissable as a solution, provided that it meets the defeat exception condition, i.e., that assumptions can be added to the set so that every model is superior to some model of the potentially defeating assumption set. Of course this extended assumption set will no longer be syntactically minimal, and hence will not be a solution to the abduction problem. However, its existence guarantees the admissibility of the original assumption set.

3 An Algorithm for Computing Abduction

Hobbs et al. [2] propose an abduction theory characterized by horn-clause rules in which antecedent literals are associated with weighting factors. I shall refer to such a theory as a weighted abduction theory; it provides a candidate for a computational realization of a model-preference abduction theory outlined int the previous section. A weighted-abduction theory is characterized by a set of literals (facts) and a set of rules expressed as implications. A general example of such a rule is

$$p_1^{w_1} \wedge \ldots \wedge p_n^{w_n} \supset q.$$

Each rule is expressed as an implication with a single consequent literal, and a conjunction of antecedent literals p_i, each associated with a weighting factor w_i. The goal of an abduction problem is expressed as a conjunction of literals, each of which is associated with an assumption cost. When proving a goal q, the abductive theorem prover can either assume the goal at the given cost, or find a rule whose consequent unifies with q, and attempt to prove the antecedent rules as subgoals. The assumption cost of each subgoal is computed by multiplying the assumption cost of the goal by the corresponding weighting factor. Each subgoal can then be either assumed at the computed assumption cost, or unified with a fact in the database (a "zero cost proof"), or unified with a literal that has already been assumed (the algorithm only charges once for each assumption instance), or another rule may be applied. The best

solution to the abduction problem is given by the set of assumptions that lead to the lowest cost proof.

A solution to an abduction problem is admissible only when all the assumptions made are consistent with each other, and with the initial theory. Therefore, a correct algorithm requires a check to filter out potential solutions that rely on inconsistent assumptions.[1]

Another possibility that must be accounted for (and which was ignored in Stickel's original formulation) is that in the frequent case in which the goal and its negation are both consistent with the theory, it will be possible to prove both the goal and its negation abductively, in the worst case by assuming them outright. This abduction algorithm guarantees that it is impossible *defeat* a proof by proving the negation of any of its assumptions at a cost that is cheaper than the cost of the proof itself.

The complete abduction algorithm can be described as follows: Given an initial theory T and a goal ϕ, generate all possible candidate assumption sets $\{A_1 \ldots A_n\}$ and sort them in order of increasing cost. Then for each successive assumption set $A_i = \{\psi_1, \ldots, \psi_m\}$, for each assumption ψ_j in A_i, attempt to prove $\neg\psi_j$ given assumptions $\psi_1, \ldots, \psi_{j-1}, \psi_{j+1}, \ldots, \psi_m$. If this proof fails (or succeeds only by assuming $\neg\psi_j$) for each j, then A_i is the best assumption set. If any $\neg\psi_j$ is provable with zero assumptions, then A_i is inconsistent and must be rejected. The remaining possibility is that $\neg\psi_j$ is provable by making some assumptions. If the cost of the best proof of any $\neg\psi_j$ is less than the cost of A_i, then A_i is defeated because its assumptions can be defeated at a lower cost than they can be assumed, and A_i is rejected in this case as well. Otherwise, A_i is contested, but not defeated, and we accept it as the best assumption set.

This algorithm can be viewed as computing solutions to an abduction problem according to the definition in the previous section, if the weighting factors on the literals can be interpreted as constraints on the model-preference relation.

A candidate interpretation of the weighting factors in terms of model preference relations is that if the weights on the antecedent literals of a rule sum to less than one, then every model that satisfies the antecedent is prefered to some model that satisfies the conjunction of the negation of the antecedent together with the consequent.

The relative magnitudes of the assumption weightings can be viewed as establishing preferences among the conclusions of different rules of the theory, provided that they obey certain constraints. If a theory contains the following two rules:

$$\begin{array}{l} p^{\alpha} \supset q \\ r^{\beta} \supset q \end{array} \quad \alpha < \beta < 1,$$

it expresses a preference for models satisfying p over those satisfying r among those models that satisfy q. Note that if r entails p, then there will be no models that satisfy $r \wedge \neg p$, and therefore, the preference relation must be circular. If the abduction algorithm were to operate on such a theory, in would incorectly compute $\{p\}$ as the best assumption set, whereas $\{r\}$ is clearly superior by the model preference criterion, because it entails p, therefore excluding every model excluded by assuming p, and other less-preferred models as well. In general, weighted abduction theories must be constrained so that the assigned weights do not imply any circularities in the model-preference relation.

4 Conclusion

The idea of characterizing domain-dependent preference among abductive assumptions as preferences among models of a theory is worthy of further investigation. What remains to be done is a full characterization of the relationship between weighted abduction and model-preference abduction, including a full specification of the relationship between rule weightings and model preferences. The incorporation of a belief operator to abstract away from particular rules of inference, following Levesque's proposal, is another interesting extension. This could lead to a knowledge-level characterization of abduction theories with domain-dependent preferences.

[1] A version of this algorithm has been implemented in the TACITUS text understanding system [2]. A version of this algorithm that is more fathful to the theory presented in this paper has been employed in plan recognition applications [1].

Acknowledgements

This research was supported by a contract with the Nippon Telegraph and Telephone Corporation. The author is grateful to David Israel and Jerry Hobbs for discussions that clarified the issues discussed herein.

References

[1] Douglas E. Appelt and Martha Pollack. Weighted Abduction as an Inference Method for Plan Recognition and Evaluation. Second International Workshop on User Modeling, proceedings forthcoming, 1990.

[2] Jerry Hobbs, Mark Stickel, Paul Martin, and Douglas Edwards. Interpretation as abduction. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, pages 95–103, 1988.

[3] Hector Levesque. A knowledge-level account of abduction. In *Proceedings of IJCAI-89*, pages 1061–1067, 1989.

[4] Judea Pearl. *Probabalistic Reasoning in Intelligent Systems*. Morgan Kaufmann, Los Altos, CA, 1988.

[5] David Poole. Explanation and prediction: an architecture for default and abductive reasoning. *Computational Intelligence*, 5(2):97–110, 1989.

[6] Raymond Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32(1):57–96, 1987.

[7] Bart Selman and Henry Kautz. The complexity of model-preference default theories. In Reinfrank et al., editor, *Non-Monotonic Reasoning*, pages 115–130, Springer Verlag, Berlin, 1989.

[8] Yoav Shoham. *Reasoning about Change: Time and Causation from the Standpoint of Artificial Intelligence*. MIT Press, Cambridge, Massachusetts, 1987.

[9] Mark E. Stickel. A prolog-like inference system for computing minimum-cost abductive explanations in natural-language interpretation. In *Proceedings of the International Computer Science Conference '88*, Hong Kong, 1988.

A Completion Semantics for Object-level Abduction

Luca Console, Daniele Theseider Dupre', Pietro Torasso

Dipartimento di Informatica
Universita` di Torino
Corso Svizzera 185 - 10149 Torino (Italy)

Introduction

Many logical formalizations of abductive reasoning have been proposed over the last few years. Most of these approaches provide meta-level definitions of the notion of abductive explanation and introduce (different) procedures to compute the causes for a set of events (see, for example, [1, 5, 7, 9, 10, 12, 13, 15]).
We believe that all these approaches are grounded on some implicit assumptions on the "abductive meaning" of a domain theory and we claim that a very clean semantics for abduction can be provided if these assumptions are made explicit. In this paper we introduce an object-level definition of abduction which is proved to be equivalent to the other definitions proposed in the literature. The object-level definition allows us to characterize abduction in a very simple way and to single out quite clearly the relationship between abduction and deduction [4].

On the Abductive Meaning of a (Causal) Theory

In order to discuss the assumptions which are implicit in most of the definitions of abduction, let us first introduce a characterization of abductive explanation which generalizes in some way those proposed in the literature: given a theory T and a formula Ψ, an explanation for Ψ in T is a set E of formulae such that
- $T \cup E$ is consistent,
- $T \cup E \vdash \Psi$,
- E has some properties that make it an interesting explanation. Typical properties are non-triviality (i.e. $E \not\vdash \Psi$) and minimality (i.e. no subset of E is an explanation).

Let us consider now a causal theory T (we assume that the cause-effect relationship between A and B is represented in the theory as the implication "$A \rightarrow B$"; the suitability of implication to model causation is beyond the scope of the paper, see [17] for an in-depth discussion) and the problem of determining an explanation for an atom "q". Suppose that the theory T contains the following formulae having "q" as their consequent:

$p_1 \rightarrow q$
$p_2 \rightarrow q$
...
$p_n \rightarrow q$
(where each p_i is a formula).

If we consider the "classical" (meta-level) definition of abduction, the process to determine the explanations for "q" (i.e. to explain why "q" is present) is based on the implicit assumption that if "q" is present, then at least one of its causes must be present and must be involved in the explanation. In particular, since "p_1", "p_2", ..., "p_n" are the only direct causes of "q" in T, then, in order for T to explain "q", at least one of them must be present.

Notice that this assumption does not mean that knowledge about "q" must be complete, but simply that one is reasoning at best of the given knowledge; i.e. it means that abduction is a defeasible form of reasoning (see [5, 6] for more comments).

If one agrees that the assumption above expresses the actual abductive meaning of a domain theory T (we shall prove that this is the case, at least as far as the definitions of abduction up to our knowledge are concerned), an interesting problem is that of looking for a syntactic transformation of T that makes such an abductive meaning explicit. Such a transformation, in fact, would provide a new semantics for abductive reasoning and would allow us to provide a clean definition of the notion of abductive explanation. The discussion above suggests that the kind of transformation that is needed is some form of circumscription (completion) of the

explainable atoms; we shall elaborate on this in the following sections.

An Object-level Account of Abduction

In this section we introduce a formal object-level account of abduction (see also [4] for a more accurate discussion). In the following, for the sake of simplicity, we shall assume that the domain theory T is a set of propositional definite clauses with acyclic dependency graph (the extension to the first-order case is straightforward and also the extension to more complex clauses requires only some more technicalities, see [3]). We shall assume, moreover, that:

- the set of predicate symbols in T is partitioned into the two disjoint subsets of the *abducible* symbols (those that can be accepted as explanations of observed data) and the *non-abducible* symbols;
- the abducible symbols are exactly those not occurring in the head of any clause[1].

Definition 1. An **abduction problem** is a pair $\langle T,\Psi \rangle$ where:
- T (the domain theory) is a set of propositional non-atomic definite clauses whose atoms are partitioned into the sets of abducible and non-abducible atoms (and whose abducible atoms are exactly those not occurring in the head of any clause);
- Ψ is a consistent conjunction of literals with no occurrence of abducible atoms (Ψ represents the observations).

Let us consider a set T of definite clauses: the *completion* [2] of non-abducible atoms in T is a set of equivalences $\{p_i \longleftrightarrow E_i \;/\; i=1,..,n\}$, where $p_1,..,p_n$ are all the non-abducible atoms (notice that on the class of theories we are considering the completion is equivalent to parallel circumscription, see [11]).

Definition 2. Let $P=\langle T,\Psi \rangle$ be an abduction problem and T_C the completion of non-abducible atoms in T. The **explanation formula** for Ψ given T is the most specific formula F in the language of abducible atoms such that:

$$T_C , \Psi \vdash F$$

where F is more specific than F' **iff** $\vdash F \rightarrow F'$.

Notice that the explanation formula characterizes all the solutions to an abduction problem P. We are interested in the most specific formula since it is the one having the highest information content among those that can be obtained from the observations Ψ. The concept of explanation formula is well defined since it can be proved from the definition that such a most specific formula exists and is unique (up to equivalence). In the following we give a procedure to determine it.

Procedure ABDUCE.
Rewrite Ψ using the equivalences in T_C (from left to right)
until a formula F containing only abducible atoms is obtained.

This procedure halts since the dependency graph of T does not contain cycles. The following correctness property can be proved[2]:

Theorem 1. Given an abduction problem $P = \langle T,\Psi \rangle$, procedure ABDUCE determines the explanation formula F for P.

We do not argue that procedure ABDUCE is the most efficient way to obtain the set of explanations: we regard it as a simple specification of such a set.

Some comments on the definitions above are worthwhile before moving to discuss the equivalence between our definition and the meta-level definitions proposed in the literature. First of all, notice that completing the non-abducible atoms in the theory T corresponds to making explicit the "abductive meaning" (abductive power) of T as discussed in the previous section. Such a completion should be automatically performed by the abductive reasoning system. The designer of a knowledge base (theory), however, should have in mind that the abductive process is based on a completion semantics, i.e. that the abductive process is based on the assumption that all the causes of each non-abducible atom are present in the model. The fact that abduction is based on a completion semantics suggests that:

[1] A complete discussion about the criteria to partition the set of symbols into the subsets of abducible and non abducible symbols is beyond the scope of the paper. We regard this problem as task or domain dependent (see also the discussion in [18]): the criteria to be adopted in causal diagnostic reasoning (where the symbols not occurring in the head of any clause are abducible) are discussed in [4, 5]; the criteria to be adopted in the planning framework are discussed in [8] while the criteria to be adopted in natural language interpretation are discussed in [18].

[2] The complete proof of this and the following theorem can be found in [3].

(1) the "plausibility" of an abductive explanation is related to the completeness of the model;

(2) one should represent explicitly the fact that some part of the model is incomplete.

In particular, the fact that not all the causes of an event A have been explicitly modeled in a theory T can be represented by adding to T a formula "$\alpha \rightarrow A$", where α is an abducible atom and denotes some unknown or unspecified cause of A (see [5, 6] for more comments on the possibility of dealing with incomplete models within abductive reasoning).

Abducible atoms are not completed since their causes are not modeled in the theory (but this does not mean that they are false as their completion would suggest).

Correspondence between the Object-level Definition and the Meta-level Ones

Let us consider now the problem of drawing the correspondence between the object-level definition of abduction presented in the previous section and the classical meta-level definitions. We shall refer, in particular, to the following meta-level definition of explanation:

Definition 3. Let $P=<T,\Psi>$ be an abduction problem, the set (conjunction) E of abducible atoms is a **m-explanation** of Ψ **iff**

(a) for every positive literal f occurring in Ψ, we have that $T \cup E \vdash f$

(b) for every negative literal $\neg f$ occurring in Ψ, we have that $T \cup E \not\vdash f$

The following theorem shows a one to one correspondence between **m-explanations** for an abduction problem P and assignments of truth values to abducible atoms satisfying the **explanation formula** for P.

Theorem 2. Let $P=<T,\Psi>$ be an abduction problem having F as the explanation formula. Let E be a set of abducible atoms and ν an assignment of truth values to the abducible atoms of T such that

$$\nu(\alpha) = \text{true} \text{ iff } \alpha \in E$$

Then E is an m-explanation for P iff $\nu \models F$.

Theorem 2 is our fundamental result showing that the meta-level definitions of abduction are indeed based on a completion semantics and highlighting the bridge between abduction and deduction through the completion semantics.

Each ν such that $\nu \models F$ can be considered as an explanation for P (notice that the case where there is no m-explanation to a problem P corresponds to the case where the explanation formula is inconsistent). If one wants to give a syntactic characterization of the notion of object-level explanation, one should consider the disjunctive normal forms of the explanation formula, as in the following definition.

Definition 4. Given an abduction problem P, each consistent disjunct of any disjunctive normal form of the explanation formula for P is an **explanation** for P.

From the practical point of view, the interesting disjunctive normal form is the minimum one (i.e. the one obtained by applying the equivalences "$X \vee \text{false} \equiv X$" and "$X \vee (X \wedge Y) \equiv X$"): it can be proved that there is a one-to-one correspondence between the positive parts of the disjuncts in the minimum disjunctive normal form of the explanation formula and the minimal (wrt set inclusion) m-explanations.

Interestingly the object-level definition we propose is more explicit than the meta-level ones in the sense that it allows us to obtain explanations in terms of both positive and negative pieces of information. Negative pieces of information are not provided explicitly by meta-level approaches (each m-explanation is a set of abducible atoms): given an m-explanation E, negative information about an abducible atom α is implicit in whether it can be consistently added to E (i.e. whether $E \cup \{\alpha\}$ is an m-explanation too). However, it could be important to distinguish clearly between

- redundant hypotheses, which appear only in non-minimal m-explanations;
- hypotheses which do not appear in *any* m-explanation, because they can be explicitly ruled out on the basis of a set of observations.

This may be particularly useful in diagnostic applications to determine the corrections (repair or therapy) to be applied to the system.

Ranking Explanations

Given an abduction problem P, in general there is more than one explanation for P. An interesting problem, therefore, is that of defining some criteria to single-out the "best explanation" to a problem or to rank alternative explanations. Many researchers suggested that such a problem is just a matter of pragmatics and that no logical criterion can be used to support the choice (see,

for example, the comments in [13, 16]). Our object level characterization, on the other hand, allows us to introduce a logical criterion (homogeneous to the definition of explanation) to compare alternative explanations and to single out the best explanation for a problem P. From the intuitive point of view, the criterion we use is the one of minimal information: we prefer those explanations which involve a minimal number of abducible literals (those that are necessary to explain the observations). The following definition characterizes which abducible atoms are necessarily true or false in a given case.

Definition 5. An abducible literal L (i.e. an abducible atom α or its negation $\neg\alpha$) is **confirmed** for an abduction problem $\langle T,\Psi\rangle$ having F as the explanation formula **iff** $F \vdash L$

This is an expression of the fact that L is necessary in order to explain the observations (see [3] and [4] for more comments and for the proof of the fact that if L is confirmed then L is a necessary assumption in order to explain Ψ). The notion of confirmation can be easily extended to explanations as follows:

Definition 6. An explanation E is **confirmed** iff $F \vdash E$

This corresponds to the case where the minimum disjunctive normal form of F has E as the unique disjunct (i.e. $F \equiv E$) and E contains only confirmed abducible literals. It should be clear that, given a problem P, there is at most one confirmed explanation for P.

The following relationships can be proved to hold (as a corollary of our main theorem):

- given the confirmed explanation, the set of its positive literals is the minimum m-explanation;
- conversely, each atom in the minimum m-explanation is confirmed.

In case the minimum disjunctive normal form of the explanation formula has more than one disjunct, then there is a similar correspondence between the positive parts of the disjuncts and the minimal (wrt set inclusion) m-explanations (as observed also in the previous section while commenting definition 4).

Notice that the correspondence is not exact since our object-level framework treats in the same way positive and negative information about abducible atoms while the criterion to prefer a minimal set of abducible atoms (intended as abnormality assumptions) is asymmetric.

In conclusion, it is important to notice that the criterion to rank explanations has been logically supported and is homogeneous with the definition of explanation.

An Example

Let us consider, as an example, a simple interpretation problem borrowed from [14]. Consider the following theory T_1:

$$T_1 = \{\ rained\_last\_night \rightarrow grass\_is\_wet,\ rained\_last\_night \rightarrow road\_is\_wet\ sprinkler\_was\_on \rightarrow grass\_is\_wet,\ grass\_is\_wet \rightarrow grass\_is\_cold\_and\_shiny,\ grass\_is\_wet \rightarrow shoes\_are\_wet\ \}$$

The atoms *rained_last_night* and *sprinkler_was_on*, which do not appear in the head of any clause, are regarded as "abducible" atoms that can be accepted as explanations of observed data.

In this case the completion gives:

$$T_1^C = \{grass\_is\_wet \longleftrightarrow rained\_last\_night \vee sprinkler\_was\_on,\ grass\_is\_cold\_and\_shiny \longleftrightarrow grass\_is\_wet,\ rained\_last\_night \longleftrightarrow road\_is\_wet\ shoes\_are\_wet \longleftrightarrow grass\_is\_wet\ \}$$

Let us consider the following abduction problem:

$P_1 = \langle T_1, \Psi_1\rangle$
where $\Psi_1 \equiv grass\_is\_cold\_and\_shiny \wedge \neg road\_is\_wet$

By applying the procedure "ABDUCE", we obtain the following explanation formula (and explanation since the formula contains only one disjunct):

$$F_1 \equiv sprinkler\_was\_on \wedge \neg rained\_last\_night$$

The example points out that our object-level approach allows us to obtain explanations in terms of both positive and negative pieces of information. Negative explanations are not provided explicitly by meta-level approaches (some approaches cannot deal at all with negative literals in the observation formula) most of which would return the explanation:

$$E_1 = \{sprinkler\_was\_on\}$$

(the negative part of the explanation is implicit in the fact that *rained_last_night* cannot be consistently added to any explanation).

Conclusions

In this paper we have introduced an object-level characterization of abduction and we have proposed a

completion semantics for abductive reasoning. Such a new characterization has several advantages since it allows us to make explicit the abductive meaning of a theory, to define abduction using very simple forms of reasoning, to obtain more explicit (precise) explanations and to introduce a logical criterion (homogeneous with the definition of explanation) to compare alternative explanations. The most interesting consequence of our definition, however, is that it highlights the relationship between abduction and deduction showing that there is a deductive solution to any abductive problem, provided that the abductive meaning of the domain theory is made explicit.

Acknowledgements. The research described in this paper has been partially supported by grants from MPI 40% (Automated Reasoning Techniques for Intelligent Systems) and CNR.

References

[1] Charniak, E., "Motivation Analysis, Abductive Unification and Nonmonotonic Equality," *Artificial Intelligence* 34(3) pp. 275-295 (1988).

[2] Clark, K., "Negation as failure," pp. 293-322 in *Logic and Data Bases*, ed. H. Gallaire, J. Minker,Plenum Press, New York (1978).

[3] Console, L., Theseider Dupre', D., and Torasso, P., "Abductive Reasoning through Direct Deduction ," Techn. Rept. Universita' di Torino (March 1989).

[4] Console, L., Theseider Dupre', D., and Torasso, P., "Abductive Reasoning through Direct Deduction from Completed Domain Models," pp. 175-182 in *Methodologies for Intelligent Systems 4*, ed. Z. Ras,North Holland (1989).

[5] Console, L., Theseider Dupre', D., and Torasso, P., "A Theory of Diagnosis for Incomplete Causal Models," pp. 1311-1317 in *Proc. 11th IJCAI*, Detroit (1989).

[6] Console, L. and Torasso, P., "Hypothetical Reasoning in Causal Models," *International Journal of Intelligent Systems* 5(1)(1990).

[7] Cox, P.T. and Pietrzykowski, T., "Causes for Events: their Computation and Application," pp. 608-621 in *Proc. Eigth Int. Conf. on Automated Deduction*, Springer Verlag Lecture Notes in Computer Science 230 (1986).

[8] Eshghi, K., "Abductive Planning with Event Calculus," pp. 562-579 in *Proceedings of the Fifth International Conference and Symposium on Logic Programming*, MIT Press, Seattle (1988).

[9] Eshghi, K. and Kowalski, R., "Abduction Compared with Negation as Failure," pp. 234-254 in *Proc. Sixth Int. Conf. on Logic Programming*, MIT Press, Lisbon (1989).

[10] Eshghi, K. and Kowalski, R., "Abduction trough Deduction," Techn. Report Imperial College of Science and Technology, University of London (1988).

[11] Genesereth, M.R. and Nilsson, N.J., *Logical Foundations of Artificial Intelligence*, Morgan Kaufmann (1987).

[12] Goebel, R., Furukawa, K., and Poole, D., "Using Definite Clauses and Integrity Constraints as the basis for a Theory Formation Approach to Diagnostic Reasoning," pp. 211-222 in *Proc. Third International Conference on Logic Programming*, Springer Verlag Lecture Notes in Computer Science 225 (1986).

[13] Levesque, H.J., "A Knowledge-level account of abduction," pp. 1061-1067 in *Proc. 11th IJCAI*, Detroit (1989).

[14] Pearl, J., "Embracing Causality in Formal Reasoning," pp. 369-373 in *Proc AAAI 87*, Seattle (1987).

[15] Poole, D., "A Logical Framework for Default Reasoning," *Artificial Intelligence* 36 pp. 27-47 (1988).

[16] Reiter, R., "Nonmonotonic reasoning," pp. 147-186 in *Annual Review of Computer Science, Vol.2*, Annual Reviews Inc. (1987).

[17] Shoham, Y., *Reasoning About Change: Time and Causation from the Standpoint of Artificial Intelligence*, MIT Press (1987).

[18] Stickel, M., "A Prolog-like Inference System for Computing Minimum-Cost Abductive Explanations in Natural Language Interpretation," Technical Note 451, Artificial Intelligence Center, SRI International, Menlo Park (1988).

Abduction and Counterfactuals

Peter Jackson

McDonnell Douglas Research Laboratories
Dept 225, Bldg 105/2, Mailcode 1065165
P.O. Box 516, St Louis, MO 63166, USA

Ginsberg (1986) showed how counterfactual logic could be used to model a kind of diagnostic inference known as 'Diagnosis From First Principles' (see *e.g.*, Genesereth, 1984; Reiter, 1987). Reiter subsequently showed that there was a close correspondence between Ginsberg's counterfactual-based method of generating diagnoses and his own consistency-based method. However, both approaches assume a complete structural description of the device under diagnosis, *i.e.*, axioms which describe all components of the device and the relationships between them. The present paper concentrates on an alternative application of counterfactual logic to diagnosis which assumes only a causal theory of the domain, *i.e.*, axioms which link cause and effect. The logic employed also corrects a number of problems with Ginsberg's original formulation. The main result is a model-theoretic demonstration of how reasoning from effects to causes (abduction) can be systematically related to belief revision using counterfactuals. Such a comparison requires precise definitions of abductive and counterfactual consequence. These are derived from the definitions in Jackson (1989a) and Jackson (1989b) respectively, each of which are primarily semantic accounts based on propositional logic.

Abductive consequence

Our notation is as follows. Letters in the range $p, q, r, \ldots$ are variables ranging over atomic formulas (atoms) of the propositional calculus (PC), while letters in the range ϕ, $\chi, \psi, \ldots$ are variables ranging over arbitrary PC formulas. Upper case letters in the range $A, B, C, \ldots$ denote sets of literals (atoms or their negations), while letters in the range $S, T, U, \ldots$ denote sets of arbitrary formulas.

Letters in the range w, v, u range over possible worlds. A possible world w will be represented by a set of atomic propositions, $[p_1, \ldots, p_n]$; the square brackets will serve to distinguish possible worlds from syntactic objects, such as sets of atoms. If p is an atom, then w satisfies p, written $w \models p$, if and only if $p \in w$, and if $p \notin w$ then $w \models \neg p$. Such worlds are essentially propositional calculus models, as specified in Chang & Keisler (1973, Ch.1). Satisfaction conditions for a compound statement ψ follow the normal truth-functional recursion on the complexity of ψ.

We assume the existence of a *causal theory*, consisting of a set of proper axioms constructed over a finite alphabet and having the general form: $fault_1 \wedge \ldots \wedge fault_m \supset symptom_1 \vee \ldots \vee symptom_n$. The axioms are used to predict how faults will manifest themselves as defective behaviour or unusual instrument readings. However, the theory may not be complete, and it will not normally amount to a full description of the device. We also assume the existence of *case data*, consisting of a finite set of literals representing such things as measurements, properties of objects (such as components) and relationships between them.

The account of propositional abduction to be found in Jackson (1989a) is most easily described in terms of model-theoretic forcing (Keisler, 1977). In the terminology of forcing, a *condition* C for a theory T is a finite set of literals consistent with T, and $C \Vdash_T Q$ denotes that C *forces* Q, *i.e.*, that $T, C \vdash Q$, where $\vdash$ denotes logical implication. In abduction, we are most interested in those minimal conditions which force the data to be true, where minimality is defined in terms of set inclusion.

Definition 1. An *explanation* E for a data set D in terms of causal theory T is a condition for T such that E forces D, *i.e.*, $E \Vdash_T D$. E is *minimal* iff there is no $E' \subset E$ such that $E' \Vdash_T D$; *non-trivial* iff $E \cap D = \emptyset$, and *less trivial than* E' iff $\{E \cap D\} \subset \{E' \cap D\}$. If E is minimal, and less trivial than any other explanation, then E is a *preferred explanation.* ||

Definition 1 is consistent with many other definitions of the term 'explanation' that can be found in the literature (*e.g.* Reggia, Nau, & Wang, 1984; Cox & Pietryzykowski, 1986). However, the formulation in terms of forcing leads naturally to a semantic account of abductive inference. In forcing theory, G is a *generic set* for T iff each $C \subseteq G$ is a condition for T and $G \Vdash_T P$ or $G \Vdash_T \neg P$ for all P. In other words, the deductive closure of $T \cup G$ is a maximal consistent set of PC formulas. As such, it characterizes a PC-model, or possible world. Given incomplete knowledge, we are most interested in sets of such worlds, and we shall call these *situations*. Intuitively, a situation for some theory is the set of all PC-models that satisfy the theory.

Definition 2. A *situation*, $\sigma_{T,D}$, for theory T and data D is a set of maximal consistent sets

$$\{\mathrm{Th}(T \cup G) \mid G \text{ is a generic set for } T \text{ that forces } D\}.$$

$\sigma_{T,D}$ is a set of models $\{w_1, ..., w_n\}$ such that $w_i \models T$ and $w_i \models D$ for all $w_i \in \sigma_{T,D}$. ||

We shall instantiate Shoham's (1988) notion of a preferential model to capture our preference for minimal, non-trivial explanations in the semantics. The following definition will be useful for characterizing preferences.

Definition 3. A set of literals E_w is a *canonical explanation* of a world $w \in \sigma_{T,D}$ iff E_w satisfies the following conditions:

(i) $w \models E_w$;
(ii) $E_w \Vdash_T F$ for all F such that $w \models F$ and $F \Vdash_T D$.

The *trivial residue*, $R_w \subseteq E_w$, of w is $E_w \cap D$. ||

The canonical explanation of a world is the smallest set of assumptions from which all explanations of D satisfied by that world follow, while the trivial residue represents the unexplained data.

Definition 4. A *preference for non-trivial, minimal explanations of D in terms of a theory T* is a strict partial order, ν, on $\sigma_{T,D}$, such that, for $w, w' \in \sigma_{T,D}$, $(w, w') \in \nu$ iff $E_w \subset E_{w'}$ and $R_w \subset R_{w'}$. The *preferred models* of $\sigma_{T,D}$ are given by the set

$$\{w \in \sigma_{T,D} \mid \neg(\exists w' \in \sigma_{T,D})(w', w) \in \nu\}.$$

A world $w \in \sigma_{T,D}$ *ν-satisfies* E_w, written $w \models_\nu E_w$, iff $w \models E_w$ and there is no $w' \in \sigma_{T,D}$ such that $w' \models E_w$ and $(w', w) \in \nu$.

The subsumption relation $E_w \subset E_{w'}$ indicates that the rival world w' overexplains the data explained by w. The subsumption relation $R_w \subset R_{w'}$ indicates that the rival w' underexplains the data, in that its canonical explanation is wholly or partly trivial. Finally, we can show that all (and only) the preferred explanations are ν-satisfiable (see Jackson, 1989a, Theorem 6).

Theorem 1. E is a preferred explanation of D in terms of T iff $w \models_\nu E$ for some $w \in \sigma_{T,D}$.

This logistic system has been dubbed PABLO, standing for 'Propositional Abductive Logic.' The following example will illustrate the method and later serve as a basis for comparison with counterfactual logic.

Example 1. Let our causal theory T be

$$\{f \supset \neg r, d \supset \neg r, f \supset \neg s, g \supset \neg s\}$$

and let our data set D be

$$\{\neg r, \neg s\}.$$

Imagine that the propositional constants in T have the following meanings: f = 'flat-battery', r = 'radio working', d = 'radio disconnected', s = 'car starts', and g = 'the car is out of gas'. Then

$$\sigma_{T,D} = \{[], [f], [d,f], [f,g], [d,g], [d,f,g]\}$$

with ν-ordering

$$[f], [d,g] > [], [d,f], [f,g], [d,f,g].$$

The preferred models are $[f]$ and $[d, g]$, representing the ν-satisfiable explanations $\{f\}$ and $\{d, g\}$.

Counterfactual consequence

The following account assumes a propositional language L, defined over a finite alphabet A. We can construct $2^{|A|}$ interpretations over this alphabet, and consider each as a possible world. Let this set of interpretations be W. A theory $S \subset L$ describes a *situation*, $W_S \in 2^W$. Thus $W_S \subset W$ is the set containing just those possible worlds in W which satisfy S.

The semantics that we shall give for counterfactuals of the form $\psi > \phi$ with respect to a theory S depends upon a very simple idea. We consider ψ: $2^W \rightarrow 2^W$ as a *revision function* that we can apply to S to return those worlds where ψ holds which are close, or most similar, to some world in W_S. $\psi > \phi$ is then a consequence of S just in case ϕ holds in each of these worlds.

Definition 5. If $B \subseteq A$, then $\Lambda_B = (2^B, \subset)$ is a *world lattice for B*. ||

Definition 6. If $\Lambda_B = (2^B, \subset)$ and $W, V \subseteq 2^B$, then $v \in V$ is a world *close to* $w \in W$, written $v \eqsim w$, iff

$$\mathbf{B}_{v,w} \vee (\exists u \in 2^B)(u \notin V \wedge \mathbf{B}_{u,w} \wedge v \eqsim u),$$

where $\mathbf{B}_{v,w}$ iff $\mathbf{glb}(v, w)$ or $\mathbf{lub}(v, w)$,
$\mathbf{glb}(v, w)$ iff $\{v' \in 2^B \mid v \subset v' \subset w\} = \emptyset$, and
$\mathbf{lub}(v, w)$ iff $\{v' \in 2^B \mid w \subset v' \subset v\} = \emptyset$.

The set of all worlds in V that are close to worlds in W, written $V \Leftarrow W$, is given by

$$V \Leftarrow W = \{v \in V \mid (\exists w \in W) v \eqsim w\}. \quad ||$$

We need to complicate the picture slightly by introducing 'bad worlds', *i.e.*, worlds which do not satisfy certain propositions that we shall deem to be protected from revision.

Definition 7. If $B \subseteq A$, and $X \subset 2^B$, then $\Lambda_{B,X} = (2^B, \subset, X)$ is a *world lattice for B w.r.t. 'bad worlds' X*. ||

Definition 8. If $\Lambda_{B,X} = (2^B, \subset, X)$ and $W, V \subseteq 2^B$, then $v \in V$ is *among the closest worlds to* $w \in W$ *avoiding worlds in X*, written $v \eqsim_X w$, iff

$$\begin{aligned}&(\mathbf{B}_{v,w} \wedge v \in (V - X)) \vee \\ &[\neg(\exists v' \in (V - X))(\mathbf{B}_{v',w}) \wedge \\ &(\exists u \in 2^B)(u \notin (V - X) \wedge \mathbf{B}_{u,w} \wedge v \eqsim_X u)].\end{aligned}$$

The *set of closest worlds in V to W avoiding X*, written $V \Leftarrow_X W$, is given by $\{v \in V \mid (\exists w \in W) v \eqsim_X w\}$. ||

Definition 9. If $S^* \subset S \subset L$, where S^* is a set of protected propositions, $\psi \in L, B \subseteq A$ contains those members of the alphabet of L occurring in either S or ψ, and $\Lambda_{B,X}$ is a world lattice for B w.r.t. $X = 2^B - W_{S^*}$, then the *semantic revision* of S by ψ w.r.t. X, written $\psi(W_S)$, is given as follows

$$\psi(W_S) = \begin{cases} W_{\{\psi\}} \cap W_S, \text{ if } W_{\{\psi\}} \cap W_S \neq \emptyset \\ \text{else } W_{\{\psi\}} \Leftarrow_X W_S. \end{cases} \quad ||$$

Definition 10. $\psi > \phi$ is a counterfactual consequence of S iff $w \models \phi$ for all $w \in \psi(W_S)$. ||

This construction has been dubbed BERYL: a failed acronym for 'Belief Revision Logic.'

It is easy to demonstrate the following result, which does not hold for Ginsberg's construction (see Jackson, 1989b, Theorem 1).

Theorem 2. If S and T are logically equivalent propositional theories and S^* and T^* are equivalent, then for all propositions ψ and ϕ, $\psi > \phi$ is a counterfactual consequence of S iff $\psi > \phi$ is a counterfactual consequence of T.

Syntax independence is obviously a good property for a belief revision function to possess. Gärdenfors (1988) identifies a number of other criteria for the classification of belief revision functions, two of which are the preservation criterion (K*P) and the monotonicity criterion (K*M). The former states that if ϕ follows from S and ψ is consistent with S, then ϕ will still follow from the revision of S by ψ. The latter states that if S and T are theories and T contains S, then the revision of T will contain the revision of S. We can show that the revisions sanctioned by BERYL are always preservative but not always monotonic.

Theorem 3. BERYL satisfies the preservation criterion: If $S \not\models \neg\psi$ and $S \models \phi$, then $\psi > \phi$.

This result holds for Ginsberg's PWA but does not hold for Winslett's (1988) 'Possible Models Approach' to belief revision, known as PMA (see Jackson, 1989b, Theorems 3 and 4). This distinction is important if one wishes to extend a belief revision system to incoporate a probabilistic model, since Bayes' Theorem endorses the preservation criterion. Thus the revision functions of BERYL and PWA are amenable to a Bayesian extension, while that of PMA is not.

BERYL is a genuinely nonmonotonic logic because, unlike PMA, it does not satisfy the monotonicity criterion (see Jackson, 1989b, Theorem 5).

Theorem 4. BERYL does not satisfy the monotonicity criterion: If $W_T \subseteq W_S$, then $\psi(W_T) \subseteq \psi(W_S)$.

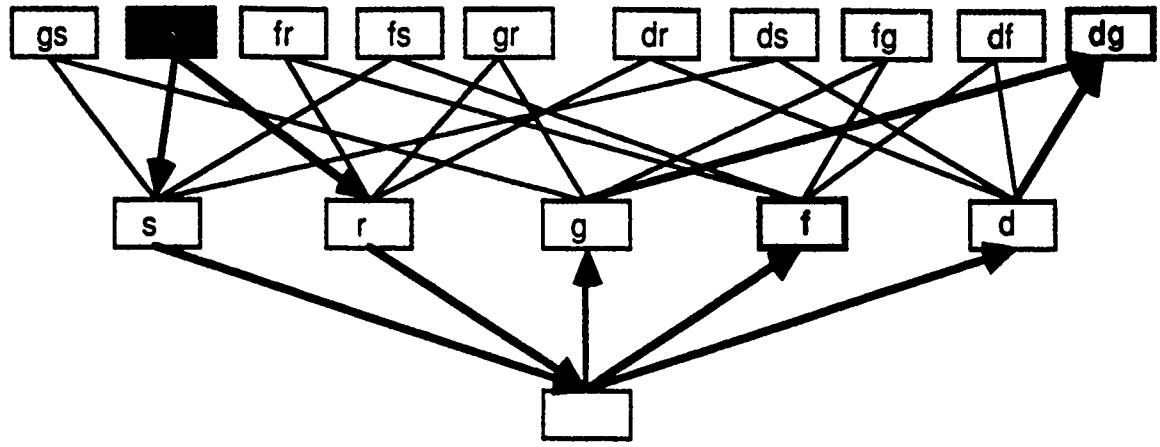


Figure 1. Part of the world lattice for Example 2. Dark shading indicates models of S; light shading indicates models of $(\neg r \wedge \neg s)$; and dark borders indicate models of $(\neg r \wedge \neg s)(W_S)$.

Finally, we redo Example 1, using counterfactual logic to perform the causal reasoning required for diagnosis. The fundamental operation involved is belief revision. We revise our original theory that the device is fault-free by the data to obtain a set of models representing diagnoses.

Example 2. First we transform the causal theory with the assumption that we know all the possible causes of the symptoms $\neg r$ and $\neg s$. We write the transformed theory, S^*, as

$$\{\neg r \supset f \vee d, \neg s \supset f \vee g\},$$

and protect every proposition in it. Then we augment S^* by assuming that all is well with the device. Thus we add the negation of all literals denoting faults or symptoms to derive

$$S = T' \cup \{r, s, \neg d, \neg f, \neg g\}.$$

These literals are not protected, however. We revise S by the data $(\neg r \wedge \neg s)$, as follows.

$$(\neg r \wedge \neg s)(W_S) = \{[f], [d, g]\}.$$

The relevant fragment of the world lattice for this problem is reproduced in Figure 1. [] is the closest world to $[r, s]$ in $W_{\{\neg r \wedge \neg s\}}$, but this world does not satisfy the protected propositions, S^*. The closest worlds to [] in $W_{\{\neg r \wedge \neg s\}}$ that satisfy S^* are $[f]$ and $[d, g]$. Note that these are precisely the preferred models of PABLO (see Example 1).

The procedure whereby we transform the causal theory T to the augmented theory T' resembles that outlined in Reiter (1987, 7.1) for effecting a 'logical reconstruction' of the GSC (Generalized Set Covering) model of Reggia, Nau, & Wang (1984). The crucial thing to notice is that this transformation amounts to the assumption that all the causes of symptoms are known. Suppose that we don't make this assumption, and attempt to revise the augmentation of the untransformed theory

$$S' = T \cup \{r, s, \neg d, \neg f, \neg g\}$$

by $(\neg r \wedge \neg s)$. Then the only model of $(\neg r \wedge \neg s)(W_{S'})$ is [], and the minimum revision of our beliefs is to attribute the symptoms to causes unknown!

Conclusions & related work

In the work of Reiter and Ginsberg, we saw that there was a connection between diagnosis from first principles using a complete structural description and counterfactual reasoning. Essentially, Reiter showed that you could do the former in terms of the latter. The present work establishes a connection between abductive reasoning from an incomplete causal theory and counterfactual logic. Again we show that the former can be done in terms of the latter, but that a particular kind of completeness assumption is required in the latter (counterfactual) treatment which is not present in the former (abductive) case. Further integration of abductive and counterfactual styles of reasoning could perhaps be achieved by a closer study of the modal foundations of conditional logic (Chellas, 1975; Segerberg, 1989).

Our counterfactual approach to causal reasoning does not require that propositions be ordered in any way, unlike Simon & Rescher's (1968) account, for example. However, this is not to say that the introduction of such orderings would not be beneficial for certain applications. The introduction of probabilities may also be highly desirable, *e.g.*, along the lines explored in Gardenfors (1988).

Acknowledgements. This work was supported by the McDonnell Douglas Independent Research and Development Program.

References

Chang, C. C. & Keisler, H. J. (1973). *Model Theory*. New York: Elsevier North Holland.

Chellas, B. (1975). Basic conditional logic. *Journal of Philosophical Logic*, 4, 133-153.

Cox, P. T. & Pietryzykowski, T. (1986). Causes for events: their computation and applications. In Sickmann, J. H. (ed.), *Proceedings of the 8th International Conference on Automated Deduction*, 608-621.

Gärdenfors, P. (1988). *Knowledge in flux*. Cambridge, MA: MIT Press.

Genesereth, M. R. (1984). The use of design descriptions in automated diagnosis. *Artificial Intelligence*, 24, 411-436.

Ginsberg, M. L. (1986). Counterfactuals. *Artificial Intelligence*, 30, 35-79.

Jackson, P. (1989a). Propositional abductive logic. In *Proceedings of the 7th Conference on Artificial Intelligence and the Simulation of Behaviour*, 89-94.

Jackson, P. (1989b). On the semantics of counterfactuals. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence*, 1382-1387.

Keisler, H. J. (1977). Fundamentals of model theory. In Barwise, J. (ed.) *Handbook of Mathematical Logic*, New York: Elsevier North-Holland.

Reggia, J. A., Nau, D. S. & Wang, P. Y. (1984). Diagnostic expert systems based on a set covering model. In Coombs, M. J. (ed.) *Developments in Expert Systems*, London: Academic Press.

Reiter, R. (1987). A theory of diagnosis from first principles. *Artificial Intelligence*, 32, 57-95.

Segerberg, K. (1989). Notes on conditional logic. *Studia Logica*, XLVIII, 157-168.

Simon, H. A. & Rescher, N. (1968). Cause and counterfactual. *Philosophy of Science*, 33, 323-340.

Shoham, Y. (1988). *Reasoning about change: Time and causation from the standpoint of artificial intelligence*. Cambridge, MA: MIT Press.

Winslett, M. (1988). Reasoning about action using a possible models approach. In *Proceedings of the 7th National Conference on Artificial Intelligence*, 89-93.

Computing Explanations

Extended Abstract

Bart Selman
Dept. of Computer Science
University of Toronto
Toronto, Canada M5S 1A4
bart@ai.toronto.edu

Abstract

Reiter and DeKleer (1987) give a precise definition of what constitutes an explanation as computed by an ATMS. We analyze the inherent computational complexity of finding such explanations.

When our underlying logical theory consists of arbitrary clauses, the task of finding any non-trivial explanation is easily shown to be NP-hard. However, when the theory contains only Horn clauses, we show that some non-trivial explanation can be found in time polynomial in the size of the theory, but that finding certain other explanations is NP-hard. We also show that the use of an assumption set in the ATMS renders the generation of an explanation computationally intractable. These results hold even for acyclic Horn theories.

Our analysis suggests that when searching for certain explanations, the method of simply listing all of them, as employed in the ATMS, cannot be improved upon. Moreover, these results show that there may not exist an appropriate restriction on the general form of the underlying theory to allow for efficient abduction. What seems to be required is some notion of an "approximate" explanation or a well-defined notion of incomplete abduction.

1 Introduction

Formal characterizations of abduction, *i.e.*, the task of finding explanations, can be divided into two camps: the set-based approaches (*e.g.*, Reggia 1983) and the logic-based ones (*e.g.*, Poole 1988). Here we will only be concerned with the latter. In particular, we will consider the type of explanation as computed by an ATMS (de Kleer 1986a; Reiter and de Kleer 1987).[1]

An ATMS computes all possible explanations for a given query. Since there may exist exponentially many such explanations, the worst-case complexity of the ATMS task is clearly exponential. However, there are situations in which one is interested in finding only one explanation or possibly a few of them. We will explore the complexity of this task.

2 Definitions

In this section, we will repeat the main definitions of the logical reconstruction of the ATMS as given by Reiter and de Kleer (1987). We will assume a standard propositional language $\mathcal{L}$. We will use p, q, r, and s (possibly with subscripts) to denote propositional letters. A clause is a disjunction of literals (a literal is either a propositional letter, called a positive literal, or its negation, called a negative literal). We will represent a clause by the set of literals contained in the clause. A clause is Horn if and only if it contains at most one positive literal. A set of Horn clauses will be called a Horn theory.

Central to Reiter and de Kleer's analysis is the notion of *prime implicant*:

Definition: Prime Implicant
A prime implicant of a set of clauses Σ is a clause C such that:

1. $\Sigma \models C$, and
2. For no proper subset C' of C does $\Sigma \models C'$.

Example: Let Σ be the set $\{\{p\}, \{q\}, \{\overline{p}, \overline{r}, s\}\}$. The prime implicants of Σ are $\{p\}$, $\{q\}$, $\{\overline{r}, s\}$, $\{r, \overline{r}\}$, and $\{s, \overline{s}\}$.

[1] The various forms of logic-based abduction are closely related (Levesque 1989).

Definition: Formal Specification of the ATMS
Given a set of Horn clauses Σ and a letter p, the ATMS procedure computes the following set:[2]

$$\mathcal{A}[\Sigma, p] = \{(q_1 \wedge \ldots \wedge q_k) \mid k \geq 0 \text{ and } \{\overline{q}_1, \ldots, \overline{q}_k, p\} \text{ is a prime implicant of } \Sigma\}$$

Since Σ together with any element from $\mathcal{A}$ logically implies p, the elements of $\mathcal{A}$ are called *explanations* for p given Σ.

Example: Let Σ again be the set $\{\{p\}, \{q\}, \{\overline{p}, \overline{r}, s\}\}$. With query s the ATMS returns the following explanations for s: r and s. We call s the trivial explanation for s; our interest lies of course in the other, non-trivial explanation.

3 Computational Complexity

Given a query p, the ATMS returns the set $\mathcal{A}[\Sigma, p]$ of all explanations for p. It is well-known that even when Σ contains only Horn clauses, there may exist exponentially many of such explanations (McAllester 1985; de Kleer 1986b). And thus, the worst case complexity of the ATMS is exponential. However, this leaves open the question of what the complexity of finding some explanation is. In particular, what is the complexity of finding a non-trivial one?

In case Σ contains arbitrary clauses, finding any non-trivial explanation is easily shown to be NP-hard.[3] However, the following theorem shows that when Σ is Horn, a non-trivial explanation (if one exists) can be computed efficiently.

Theorem 1 *Given a set of Horn clauses Σ and a letter p, a non-trivial explanation for p can be computed in time $O(kn)$, where k is the number of propositional letters and n is the number of occurrences of literals in Σ.*

Here we only give an outline of the algorithm. Consider a Horn clause in Σ of the following form: $\{\overline{q}_1, \ldots, \overline{q}_k, p\}$ with $k \geq 0$ (if no such clause exists, return "no non-trivial explanation"). Now, clearly $\alpha = (q_1 \wedge \ldots \wedge q_k)$ with Σ implies p. Subsequently, try removing a letter from α while ensuring that the remaining conjunction together with Σ implies p (testing can be done in linear time, using the procedure by Dowling and Gallier (1984)). Repeat this process until no more letters can be removed. If the remaining conjunction is non-empty and combined with Σ is consistent, return that one; otherwise consider another clause containing p and repeat the above procedure. When all clauses containing p have been explore and no explanation is found, return "no non-trivial explanation."

From the algorithm, it is clear that we find only certain, very particular explanations — which ones will strongly depend on the particular form in which the background knowledge Σ is written down. This raises the question whether there is an efficient procedure to generate other explanations. In particular, suppose one is interested in only those explanations for p that contain a given set of letters S, can one efficiently find such explanations? The set S could be used, for example, to identify components that have a high failure rate when doing circuit diagnosis. We will term this form of reasoning *goal-directed* abduction.

Note that the notion of goal-directed abduction is especially relevant in light of the fact that there are often exponentially many explanations; simply listing them all would be prohibitive. One therefore has to be selective in generating explanations. Goal-directed abduction allows one to consider only certain subsets of explanations that are of particular interest.[4] But what is the computational cost of such reasoning?

Unfortunately, the following theorem shows that there is no efficient algorithm for goal-directed abduction.

Theorem 2 *Given a set of Horn clauses Σ, a letter p, and a set of letters S, the problem of generating an explanation for p that contains the letters from S is NP-hard.*

The proof of this theorem is based on a reduction from the NP-complete decision problem "path with forbidden pairs" defined by Gabow, Maheshwari, and Osterweil (1976) (see also Garey and Johnson 1979). We will give the details of this reduction in the full paper.

Intuitively speaking, theorem 2 shows that certain explanations will be hard to find, even if our background theory Σ is Horn. This result also holds when S contains

[2] De Kleer identifies a subset A of the set P of propositional letters in Σ, and requires that each element in $\mathcal{A}[\Sigma, p]$ contains only elements from A. For now, we will assume that $A = P$. This does not affect the notion of explanation. We will return to this issue later on.

[3] We haven't actually defined the notion of explanation given a non-Horn theory. However, the definition follows from a straightforward generalization of the ATMS, in which we simply allow arbitrary clauses in Σ. Since non-trivial explanations can only exist when Σ is consistent, satisfiability testing of CNF formulas can be reduced to the problem of finding a non-trivial explanation.

[4] For a related approach, see De Kleer and Williams (1989).

only a single letter and when Σ consists of an acyclic Horn theory.[5]

Finally, we consider the influence of the use of an *assumption set* in the ATMS. An assumption set A is a distinguished subset of the propositional letters in Σ. Given a query p, the ATMS will generate only explanations that contain letters from among those in A. Note that the assumption set again allows one to select a certain subset of all possible explanations. An assumption may represent a hypothesis that one is willing to consider as part of an explanation of the query.[6] The following theorem shows that the use of such an assumption set dramatically increases the complexity of finding a non-trivial explanation (compare with theorem 1):

Theorem 3 *Given a set of Horn clauses Σ, a letter p, and a set of assumptions $A \subseteq P$, finding a non-trivial explanation for p with letters among A is NP-hard*

The proof of this theorem is based on a modification of the reduction used in the proof of theorem 2 (details will be given in the full version of the paper).

This theorem shows that apart from the fact that the ATMS (with $A \subseteq P$) may have to list an exponential number of explanations, merely finding one of them may require exponential time.

Our reductions are based on a theory with exponentially many prime implicants for the query p. Therefore, our analysis does not give us the complexity of generating explanations given a background theory that has only polynomially many prime implicants containing p. Since such theories may have practical significance, we are currently investigating the complexity of such theories.

4 Conclusions

We have shown that given a Horn theory and a letter p, some non-trivial explanation for p can be computed in polynomial time. However, goal-directed abduction or the use of an assumption set renders the problem of computing an explanation intractable, even for acyclic Horn theories. Thus, the exponential worst-case complexity of the ATMS is not just a consequence of the fact that the ATMS may have to list an exponential number of explanations; even if we insist only on the generation of one or a few explanations that contain letters from the given assumption set, the task remains inherently intractable. Thus, it appears unlikely that the efficiency of the ATMS algorithm can be significantly improved.

Our results show that abduction is inherently hard. In fact, there may not exist appropriate restrictions on the general form of the underlying theory to allow for efficient abduction. This situation should be contrasted with that for deductive and default reasoning: there exists a linear time procedure for dealing with propositional Horn theories (Dowling and Gallier 1984); and there are polynomial algorithms for dealing with certain acyclic default theories (Selman and Kautz 1988; Kautz and Selman 1989). What seems to be required is some notion of an "approximate" explanation or a well-defined notion of incomplete abduction (a proposal for the latter is given in Levesque 1989).

[5] Given a Horn theory Σ, let G be a directed graph containing a vertex for each letter in Σ and an edge from any vertex corresponding to a letter on the left-hand side of a Horn rule to the vertex corresponding to the letter on the right hand-side of that rule. A Horn theory is acyclic if and only if the associated graph G is acyclic.

[6] This way of selecting certain explanations is closely related, but not identical, to the notion of goal-directed abduction.

Acknowledgments

I would like to thank Hector Levesque for introducing to me the problem of computing explanations based on Horn theories, and Ray Reiter for useful discussions and comments.

References

de Kleer, J. (1986a) An assumption-based TMS, *Artificial Intelligence*, Vol. 28, #2, 1986, 127–162.

de Kleer, J. (1986b) Problem solving with the ATMS, *Artificial Intelligence*, Vol. 28, #2, 1986, 197–224.

de Kleer, J. and Williams, B.C. (1989) Diagnosis with behavioral modes. *IJCAI-89*, Detroit, MI, 1325-1330.

Dowling, W.F. and Gallier, J.H. (1984). Linear-Time Algorithms for Testing the Satisfiability of Propositional Horn Formulae. *Journal of Logic Programming*, **3**, 1984, 267–284.

Gabow, H.N., Maheshwari S.N, and Osterweil L. (1976). On Two Problems in the Generation of Program Test Paths. *IEEE Trans. Software Engineering*, 1976, 227–231.

Garey, M.R. and Johnson, D.S. (1979). *Computers and Intractability, A Guide to the Theory of NP-Completeness.* New York: W.H. Freeman, 1979.

Kautz, Henry A., and Selman, Bart. (1989) Hard Problems for Simple Default Logics, *Proceedings of*

the First International Conference on Knowledge Representation and Reasoning (KR-89), Toronto, Ont., 189–197.

Levesque, H.J. (1986). A knowledge-level account of abduction. *IJCAI-89*, Detroit, MI, 1061–1067.

McAllester, D. (1985). A widely used truth maintenance system. MIT, AI-lab memo, 1985.

Poole, D. (1988). A methodology for using a default and abductive reasoning system. Technical report, Dept. of Computer Science, University of Waterloo, Waterloo, Ont., 1988.

Reggia, J. (1983) Diagnostic expert systems based on a set-covering model. *International Journal of Man Machine Studies*, 19(5), 1983, 437–460.

Reiter, R. and de Kleer, J. (1980). Foundations of assumption-based truth maintenance systems. *AAAI-87*, Seattle, WA, 1987, 183–188.

Selman, Bart and Kautz, Henry A. (1988) The Complexity of Model-Preference Default Theories. *Proceedings of the Seventh Biennial Conference of the Canadian Society for Computational Studies of Intelligence (CSCSI-88)*, Edmonton, Alberta, 102–109. Extended version to appear in *Artificial Intelligence*.

When Efficient Assembly Performs Correct Abduction and Why Abduction Is Otherwise Trivial or Intractable

Tom Bylander

Laboratory for Artificial Intelligence Research
Department of Computer and Information Science
The Ohio State University
Columbus, Ohio

Introduction

We have formally analyzed several classes of abduction problems, which fall into the following three categories: (1) intractable (NP-hard) because of combinatorial interactions among hypotheses; (2) "trivially" tractable because a polynomial hypothesis space is guaranteed; or (3) tractable by hypothesis assembly (Josephson et al., 1987) within a "nontrivial" hypothesis space. While our analysis does not exhaust all possible classes of abduction problems, it strongly suggests that efficient and effective abduction in nontrivial domains is possible only by satisfying the constraints required for hypothesis assembly.

In this extended abstract, we describe our model of abduction and the above classes of problems. Considerably more detail and the historical progression of our analysis can be found in Allemang et al. (1987) and Bylander et al. (1989a; 1989b).

The Model

Our model of abduction characterizes the abductive task as finding the most plausible explanation of a set of data. We use the following notational conventions and definitions. d will stand for a datum, e.g., a symptom. D will stand for a set of data. h will stand for an individual hypothesis, e.g., a hypothesized disease. H will stand for a set of hypotheses, which can itself be considered a composite hypothesis, e.g., an hypothesized set of diseases.

An *abduction problem* is a tuple $\langle D_{all}, H_{all}, e, pl \rangle$, where:

D_{all} is a finite set of all the data to be explained,

H_{all} is a finite set of all the individual hypotheses,

e is a map from subsets of H_{all} to subsets of D_{all} (H explains $e(H)$), and

pl is a map from subsets of H_{all} to a partially ordered set (H has plausibility $pl(H)$).

We require that pl satisfies a weak version of Occam's Razor:

$$\forall H, H' \in H_{all}(H \subseteq H' \rightarrow pl(H) \geq pl(H'))$$

That is, a composite hypothesis cannot be more plausible than any of its subsets.

H is *complete* iff:

$$e(H) = D_{all}$$

That is, H explains all the data.

H is *parsimonious* iff:

$$\nexists H' \subset H(e(H) \subseteq e(H'))$$

That is, no proper subset of H explains what H does.

H is an *explanation* iff

$$\text{complete}(H) \wedge \text{parsimonious}(H)$$

That is, H explains all the data and has no superfluous elements.

H is a *best explanation* iff:

$$\text{explanation}(H) \wedge \nexists H' \subseteq H_{all}\,(\text{explanation}(H') \wedge pl(H') > pl(H))$$

In other words, no other explanation has a higher plausibility than H. Because the range of pl is permitted to be a partial ordering, there might be several "best explanations."

Our model leaves e and pl virtually unconstrained. We exploit this freedom below by defining and analyzing natural constraints on e and pl without considering the representations—logical, causal, or probabilistic—underlying the computation of e and pl.

We do assume, though, the tractability of e and pl, as well as an "inverse" function, denoted as e^{-1}, from D_{all} to subsets of H_{all}, defined as:

$$e^{-1}(d) = \{h |\ \exists H \subset H_{all}(d \notin e(H) \wedge d \in e(H \cup \{h\}))\}$$

That is, $e^{-1}(d)$ is the set of individual hypotheses that can contribute to explaining d. We denote the time

complexity of e, e^{-1}, and pl with respect to the size of an abduction problem as C_e, $C_{e^{-1}}$, C_{pl}, respectively. These functions are not necessarily tractable (Pearl, 1987; Reiter, 1987), but making the assumptions simplifies our analysis.

These definitions and assumptions simplify several aspects of abduction. For example, we define composite hypotheses as conjunctions of individual hypotheses. In reality, the relationships between the parts of an abductive answer can be much more complex, both logically and causally. Despite this and other simplifications, our analysis provides powerful insights concerning the computational complexity of abduction.

Using our model of abduction, we discuss, in order, trivial classes of abduction problems, intractable classes, and tractable, but nontrivial, classes.

The Trivial

A *trivial* class of abduction problems is when some constraint guarantees a polynomial hypothesis space. There are several ways this can occur.

The single fault constraint. If the individual hypotheses are mutually exclusive, then no multi-part hypotheses need to be considered. More generally, if the size of composite hypotheses is limited by a constant k, then there are $O(n^k)$ composite hypotheses.

The rule-out constraint. Let pl_0 be the lowest possible plausibility value. Let $H = \{h|\ pl(h) = pl_0\}$, i.e., H is the set of individual hypotheses that are ruled out from consideration in any composite hypothesis. If all but $O(\log n)$ individual hypotheses are ruled out, then only $O(n)$ composite hypotheses need to be considered.

The pathognomonic constraint. A datum might have only one individual hypothesis that can explain it, i.e., the datum is pathognomonic for that hypothesis. Let $H = \{h|\ \exists d \in D_{all}\ (e^{-1}(d) = \{h\})\}$. That is, H is the set of hypotheses that are indicated by pathognomonic data. If $e(H) = D_{all}$, then H is the only explanation. More generally, if only $O(\log n)$ individual hypotheses can contribute to explaining the remaining data $D_{all} \setminus e(H)$, then only $O(n)$ composite hypotheses need to be considered (assuming no cancellation effects).

The pathognomonic-after-rule-out constraint. A datum might have only one plausible individual hypothesis that can explain it, i.e., the other individual hypotheses explaining the datum are ruled out. The same analysis as in the previous paragraph applies if only non-ruled-out individual hypotheses are considered.

These constraints trivialize complexity analysis because exhaustive search over the possible composite hypotheses becomes a tractable strategy, hence the label "trivial." Trivial does not imply unimportant, though. To the contrary, knowledge engineering of abductive systems should try to satisfy, if possible, the above constraints because they guarantee efficient and effective abduction. However, for complex domains, it is unlikely that such knowledge can be engineered.

The Intractable

First, we shall consider different constraints on e, and then consider properties of pl.

Independent Abduction Problems

In the simplest problems, an individual hypothesis explains a specific set of data regardless of what other individual hypotheses are being considered. Formally, an abduction problem is *independent* iff:

$$\forall H \subseteq H_{all}\ (e(H) = \bigcup_{h \in H} e(h))$$

That is, a composite hypothesis explains a datum if and only if one of its elements explains the datum. This constraint makes explanatory coverage equivalent to set covering (Reggia et al., 1983). Many other abduction approaches make similar assumptions (de Kleer and Williams, 1987; Eshelman et al., 1987; Miller et al., 1982; Pearl, 1987; Peng and Reggia, 1987; Reiter, 1987).

One way to find a best explanation would be to generate all explanations and then sort them by plausibility. Unfortunately, *for the class of independent abduction problems, generating all explanations is NP-hard.* If this task were tractable, then the minimal set cover problem (Garey and Johnson, 1979) could be tractably solved by first generating all parsimonious set covers and then selecting the smallest one. But minimal set cover is known to be NP-hard; thus, so is generating all explanations. However, the definition of best explanation does not require that all explanations be explicitly enumerated, a fact that we shall later rely on.

Monotonic Abduction Problems

A more general kind of problem is when a composite hypothesis can explain additional data that are not explained by any of its elements. Formally, an abduction problem is *monotonic* iff:

$$\forall H, H' \subseteq H_{all}\ (H \subseteq H' \rightarrow e(H) \subseteq e(H'))$$

Since the class of monotonic abduction problems is a superset of independent ones, *for the class of monotonic abduction problems, generating all explanations is NP-hard.*

Incompatibility Abduction Problems

So far we have assumed that any collection of individual hypotheses is possible. In general, however, the negation of a hypothesis can also be considered a hypothesis.

Formally, an *incompatibility abduction problem* is a tuple $\langle D_{all}, H_{all}, e, pl, \mathcal{I} \rangle$, where D_{all}, H_{all}, e, and pl are the same as before and $\mathcal{I}$ is a set of two-element subsets of H_{all}, indicating pairs of hypotheses that are incompatible with each other. No composite hypothesis containing an incompatible pair can be an explanation.

Even if an abduction problem is otherwise independent, it is difficult to even find a single explanation. *For the class of incompatibility abduction problems, determining whether an explanation exists is NP-complete.* This is because it can be difficult to choose between incompatible hypotheses. Only $O(n)$ incompatible pairs are needed for this result.

Cancellation Abduction Problems

Another interaction not allowed in independent or monotonic abduction problems is cancellation, i.e., when one element of a composite hypothesis "cancels" a datum that another element would otherwise explain. Cancellation can occur when one hypothesis can have a subtractive effect on another (Patil et al., 1982).

Formally, we define a *cancellation abduction problem* as a tuple $\langle D_{all}, H_{all}, e, pl, e_+, e_- \rangle$. e_+ is a map from H_{all} to subsets of D_{all} indicating what data each hypothesis "produces." e_- is another map from H_{all} to subsets of D_{all} indicating what data each hypothesis "consumes." $d \in e(H)$ iff the number of hypotheses in H that produce d outnumber the hypotheses that consume d. That is:

$$d \in e(H) \leftrightarrow |\{h |\ h \in H \wedge d \in e_+(h)\}| > |\{h |\ h \in H \wedge d \in e_-(h)\}|$$

Admittedly, this is an oversimplified model of cancellation effects, in the sense that it captures only one kind of cancellation interaction. Nevertheless, *for the class of cancellation abduction problems, it is NP-complete to determine whether an explanation exists.* Even if a complete composite hypothesis is found, *for the class of cancellation abduction problems, it is NP-complete to determine whether a complete composite hypothesis is not parsimonious.* The difficulty arises when several data each has several consumers. Either a difficult choice between the consumers of each datum must be made, or sufficient producers must be added for each datum, possibly violating parsimony.

The Best-Small Plausibility Criterion

Clearly, finding a best explanation for incompatibility and cancellation abduction problems is NP-hard. To determine the complexity of finding a best explanation in independent and monotonic abduction problems, properties of plausibility must be analyzed.

Everything else being equal, smaller explanations are preferable to larger ones, and more plausible individual hypotheses are preferable to less plausible ones. Thus, in the absence of other information, it is reasonable to compare explanations based on their sizes and the relative plausibilities of their elements.

The *best-small plausibility criterion* formally characterizes these considerations as follows:

$$\begin{aligned} & pl(H) > pl(H') \leftrightarrow \\ & \exists m\colon H \rightarrow H' \,(m \text{ is 1-1} \wedge \\ & \qquad \forall h \in H\ (pl(h) \geq pl(m(h))) \wedge \\ & \qquad (|H| = |H'| \rightarrow \\ & \qquad\ \exists h \in H\ (pl(h) > pl(m(h))))) \end{aligned}$$

That is, to be more plausible according to best-small, the elements of H need to be matched to the elements of H' so that the elements of H are at least as plausible as their matches in H'. If H and H' are the same size, then in addition some element in H must be more plausible than its match in H'. Note that $|H| > |H'| \rightarrow pl(H) \not> pl(H')$.

Even for independent abduction problems, it is difficult to find any best explanations according to best-small. *For the class of independent abduction problems using the best-small plausibility criterion, given an explanation, it is NP-complete to determine whether a better explanation exists.* This class of problems is hard because it is difficult to choose between individual hypotheses with equal or similar plausibility.

The Tractable

What if the individual hypotheses all have different plausibilities? We call such an abduction problem *ordered*, formally defined as:

$$\forall h, h' \in H_{all}\ (h \neq h' \rightarrow (pl(h) < pl(h') \vee pl(h) > pl(h')))$$

It turns out that this condition permits tractable abduction. *For the class of ordered independent abduction problems using the best-small plausibility criterion, there is an $O(nC_e + nC_{e-1} + nC_{pl} + n^2)$ algorithm for finding a best explanation.* Algorithm 1 performs this task within this order of complexity. The explanation found by this algorithm is a best explanation because it always chooses the most plausible individual hypotheses

W stands for the working composite hypothesis.
Nil is returned if no explanation exists.

Find a complete composite hypothesis.
More plausible individual hypothesis are preferred.
$W \leftarrow \emptyset$
For each $d \in D_{all}$
    If $e^{-1}(d) = \emptyset$ then
        Return nil
    else $W \leftarrow W \cup \{h\}$ such that
        $h \in e^{-1}(d) \wedge$
        $\forall h' \in e^{-1}(d)\ (pl(h) \geq pl(h'))$

Find a parsimonious subset.
Try to remove less plausible hypotheses first.
For each $h \in W$ from least to most plausible
    If $e(W \backslash \{h\}) = e(W)$ then
        $W \leftarrow W \backslash \{h\}$
Return W

Algorithm 1: Finding a Best Explanation in Ordered Independent Abduction Problems

to keep, and the least plausible individual hypotheses to remove. This algorithm is a variant of the hypothesis assembly strategy described in Josephson et al. (1987).

Because best-small in general imposes a partial ordering on the plausibilities of composite hypotheses, there might be more than one explanation. Suppose that an ordered independent abduction problem had only one best explanation according to best-small. Because Algorithm 1 is guaranteed to find a best explanation, then it will find the one best explanation. *For the class of ordered independent abduction problems using the best-small plausibility criterion, if there is exactly one best explanation, then Algorithm 1 finds the best explanation.* This can be informally restated as: *In a well-behaved abduction problem, if it is known that some explanation is clearly better than any other explanation, then it is tractable to find it.*

Unfortunately, it is difficult to determine if there is exactly one best explanation. *For the class of ordered independent abduction problems using the best-small plausibility criterion, it is NP-complete to determine whether there is more than one best explanation.* The reason for the difficulty is that any other best explanations will be smaller than the one found by Algorithm 1. Consequently, determining that no other best explanation exists is equivalent to determining that the explanation found by the algorithm is a minimal set cover. Thus, even for ordered independent abduction problems, it is intractable to generate all the best explanations.

From these results, we can describe what kinds of mistakes will be made by Algorithm 1. While the explanation it finds will be able to match up qualitatively to any other explanation, there might be smaller explanations that are better based on quantitative information. Similar results hold for ordered monotonic abduction problems.

Conclusion

Based on these results, we propose that one of the following properties must be satisfied for abduction to be effective and efficient.

> The domain is trivial in the sense described above. In other words, sufficient knowledge exists to engineer abduction problems so that exhaustive search is tractable, i.e., by selecting appropriate data and hypotheses and by constructing powerful e and pl functions.
>
> The domain satisfies the monotonic and ordered properties, and there exists one best explanation according to best-small. This, too, might call for considerable knowledge engineering. Hypothesis assembly is directly applicable to such domains.
>
> Incompatibility relationships, cancellation interactions, and unordered hypotheses must be sparse ($\leq O(\log n)$), and otherwise the domain satisfies the monotonic, ordered, and one-best-explanation properties. In these domains, the best explanation can be found by invoking hypothesis assembly a polynomial number of times, i.e., by varying the choices from incompatible pairs of hypotheses, consumers of each datum, and unordered hypotheses.

Of course, if more than one property can be satisfied, so much the better.

Our analysis supports the following thesis: *if hypothesis assembly cannot be used to find the best explanation in a complex domain, then the domain is intractable.*

Acknowledgments. Thanks to Dean Allemang, Mike Tanner, and John Josephson for their comments. This research has been supported by the National Heart, Lung and Blood Institute, NIH Grant 1 R01 HL 38776-01.

References

Allemang, D., Tanner, M. C., Bylander, T., and Josephson, J. R. (1987). On the computational complexity of hypothesis assembly. In *Proc. Tenth Int. Joint Conf. on Artificial Intelligence*, pages 1112–1117, Milan.

Bylander, T., Allemang, D., Tanner, M. C., and Josephson, J. R. (1989a). Some results concerning the computational complexity of abduction. In *Proc. First Int. Conf. on Principles of Knowledge Representation and Reasoning*, pages 44–54, Toronto.

Bylander, T., Allemang, D., Tanner, M. C., and Josephson, J. R. (1989b). The computational complexity of abduction. Technical report, Lab. for AI Research, CIS Dept., Ohio State Univ., Columbus, OH.

de Kleer, J. and Williams, B. C. (1987). Diagnosing multiple faults. *Artificial Intelligence*, 32(1):97–130.

Eshelman, L., Ehret, D., McDermott, J., and Tan, M. (1987). MOLE: A tenacious knowledge acquisition tool. *Int. J. Man-Machine Studies*, 26(1):41–54.

Garey, M. R. and Johnson, D. S. (1979). *Computers and Intractability*. W. H. Freeman, New York.

Josephson, J. R., Chandrasekaran, B., Smith, J. W., and Tanner, M. C. (1987). A mechanism for forming composite explanatory hypotheses. *IEEE Trans. Systems, Man, and Cybernetics*, 17(3):445–454.

Miller, R. A., Pople, J. E., and Myers, J. D. (1982). INTERNIST-I, An experimental computer-based diagnostic consultant for general internal medicine. *New England J. of Medicine*, 307:468–476.

Patil, R. S., Szolovits, P., and Schwartz, W. B. (1982). Modeling knowledge of the patient in acid-base and electrolyte disorders. In Szolovits, P., editor, *Artificial Intelligence in Medicine*, pages 191–226. Westview Press, Boulder, CO.

Pearl, J. (1987). Distributed revision of composite beliefs. *Artificial Intelligence*, 33(2):173–215.

Peng, Y. and Reggia, J. A. (1987). A probabilistic causal model for diagnostic problem solving. *IEEE Trans. on Systems, Man, and Cybernetics*, 17(2):146–162. Part II appears in SMC-17(3):395–406.

Reggia, J. A., Nau, D. S., and Wang, P. (1983). Diagnostic expert systems based on a set covering model. *Int. J. Man-Machine Studies*, 19:437–460.

Reiter, R. (1987). A theory of diagnosis from first principles. *Artificial Intelligence*, 32(1):57–95.

Towards Taxonomies of Abductive Systems at the Knowledge and Symbol Level

Jack Smith, Jr. M.D.,Ph.D. and Olivier Fischer
Laboratory for Knowledge Based Medical Systems
Department of Pathology
The Ohio State University
Columbus, Ohio 43210
Netmail:jsmith@cis.ohio-state.edu

Introduction

In this abstract we will summarize some of the results of our knowledge and symbol level analysis of two medical AI systems: RED-2 [Josephson, et. al.1987] [Smith, et.al. 1986] and INTERNIST-1 [Miller, et. al. 1982], [Pople, et. al.., 1975] , [Pople, 1977]. RED-2 and INTERNIST-1 were both intended to encode an abductive problem-solving method for different kinds of domain problems and with different problem-solving methods.

For this analysis we adopted Newell's view of what a knowledge vs. symbol level analysis should address [Newell, 1981]. In this view, a knowledge level analysis of a system should focus on a representation independent description of the system's goals and the bodies of knowledge it brings to bear to satisfy these goals. At the symbol level, the analysis should center on the representations that attempt to realize this knowledge level. For this purpose we adopted a similar analysis methodology and set of terms to Clancey [Clancey, 1985]. In particular, we use his taxonomy of kinds of problems and his method of specifying the content of a knowledge base by describing the methods and knowledge applied. We also borrow the concept types and conceptual relations he used in describing the heuristic classification method.

Clancey performed such an analysis on a wide variety of rule-based systems with interesting results. Similarly, interesting similarities and differences emerged in our analysis when the two systems we studied were described using the same framework and a common vocabulary.

Knowledge Level Analysis of Two Abductive Systems

We will now individually describe the two systems in terms of the type of domain problem they solve and their problem solving methods. We describe their methods in terms of their abstract goals, subgoals, and the kinds of knowledge they attempt to apply to achieve them. We will then compare the two systems along several dimensions. We argue that they encode in different forms very similar goals, methods, and kinds of knowledge. We then suggest that general characteristics of their respective domain problems and similar assumptions underlying their methods are the reason why the designers of RED-2 and INTERNIST-1 made similar content but different form decisions.

INTERNIST-1

Following Clancey, the problems INTERNIST-1 tries to solve are diagnostic. In his problem taxonomy diagnosis problems are a kind of interpretation problem requiring the identification as faulty some part of the system that is being diagnosed with respect to a preferred model of the system. Interpretation is concerned with making assertions about a working system in some environment while identification requires taking descriptions of input/output behavior and mapping it onto a system. More specifically, a solution to INTERNIST-1's problems is a combination of diseases(faults of the human body) which are the cause of the patient's manifestations (data). INTERNIST-1 is designed to search for these diseases using a method which selects the best explanation found for the manifestations using the subgoals shown in Figure 1. The two top subgoals are to select the diseases(solutions) which explain all

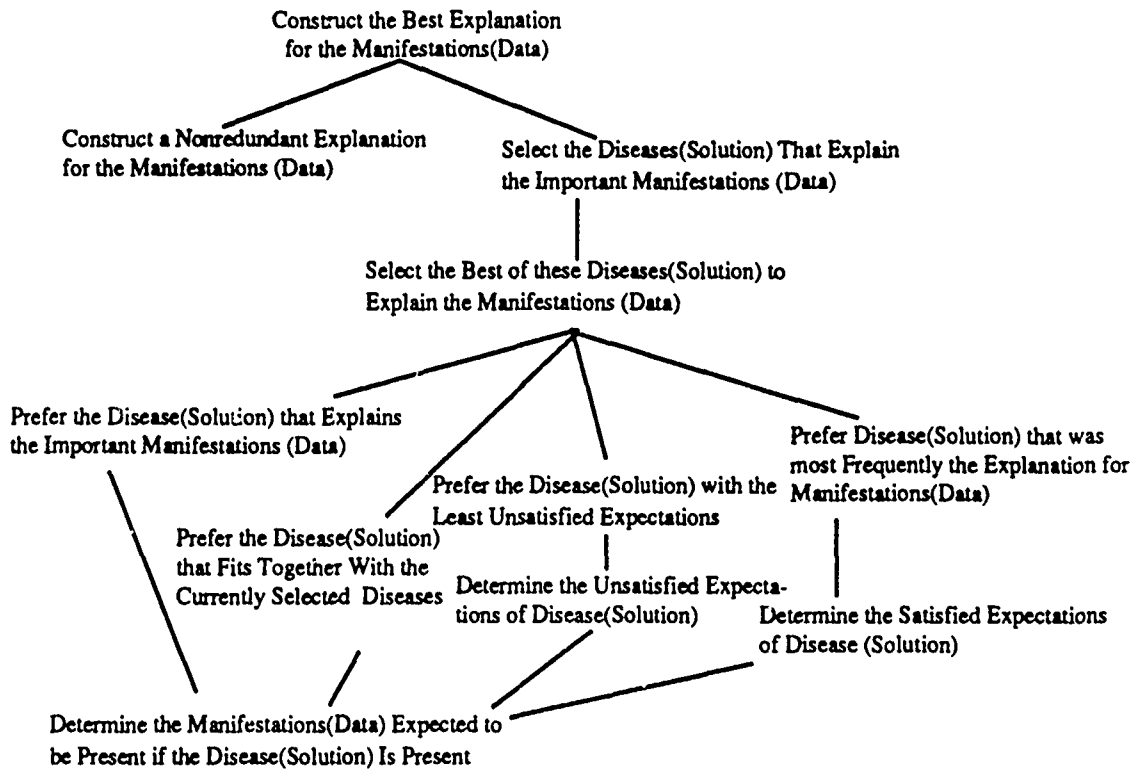


Figure 1. Problem-Solving Method Goal/Subgoal Structure in INTERNIST-1

the important manifestations(data) and construct a nonredundant explanation.

Determining that all the important mani- festations are explained is achieved by encoding for every manifestation a qualitative importance of it being explained by the final diagnosis. INTERNIST-1 will attempt this goal until all the important data above a certain importance value is explained. The goal of explaining the important manifestations is decomposed into a series of instances of the goal of selecting the best individual disease to explain the manifestations that remain unexplained. This goal decomposition is justified assuming that the causal processes underlying the diseases have minimal interactions . More precisely, the designers of INTERNIST-1 explicitly assume the diseases defined in the system are mutually exclusive causes of any datum. This allows for explaining the manifestations by performing a succession of single-disease diagnoses.

The goal of choosing the best single disease is then decomposed into four subgoals to prefer the single disease that: explains the remaining manifestations best, most frequently was the best explanation for the manifestations in the past (in their terminology has the highest evoking strength), that fits best with the currently selected diseases, and has the least unsatisfied expectations. The combined degrees of satisfaction of a disease to these criteria determines the degree of satisfaction of the individual disease to explain the data.

INTERNIST-1 explicitly encodes for every individual manifestation qualitative knowledge regarding the frequency with which disease(s) are found to be the best explanation for the manifestation. This knowledge is used to prefer the disease with the highest value for the manifestations at hand. The goal of selecting the disease that best explains the manifestations is achieved by selecting the disease whose expected manifestations match the most important manifestations in the case at hand. Also, for each disease INTERNIST-1 has knowledge encoded regarding the frequency with which a manifestation occurs if the disease is present. Selecting the disease with the least unsatisfied expectations can be accomplished by selecting the disease whose expected manifestations are the most completely matched by the case at hand. Knowledge is encoded to allow for diseases to be preferred that are known to predispose or be causally related to another diseases that has already been determined to be part of the solution. This knowledge is used to determine the

degree of fit of a disease to the current solution.

The top-level goal of constructing a non-redundant final diagnosis is achieved partly by discarding the manifestations related to diseases already included in what will become the final diagnosis. This is consistent with the assumption of exhaustivity and mutual exclusivity of the diseases which we have already noted as a basic premise of INTERNIST-1. Therefore the knowledge applied to avoid redundant explanations in the final diagnosis is knowledge indicating what manifestations each disease can account for and the knowledge of what diseases have already been established as parts of the final solution. In addition, there is the assumption that once a disease is selected as part of the solution to the problem, the manifestations it explains do not need an explanation anymore and cannot help in selecting other diseases.

RED-2

By contrast, RED-2 does not solve medical diagnosis problems but a medical identification problem. RED-2 solves problems which involve interpreting laboratory tests to identify red cell antibodies in the serum of patients in order to safely transfuse red cells. The presence of these antibodies do not indicate malfunctions of the human body. A combination of antibodies which are the cause of the test reactions is considered a solution to this problem. Using Clancey's terminology, the systems to be described are the test tubes where the potential recipient's serum is mixed with the other components of the test system(like standardized red cells). Abstractly, the problem involves taking descriptions of the output behavior of the test system in the form of test reactions and mapping it back onto the test system by determining what antibodies in the serum give rise to the reactions.

The problem-solving method RED-2 is designed to apply is shown in Figure 2. To select the best explanation found during this search, RED-2 prefers combinations of antibodies(solution) which non-redundantly explain all the test reactions(data) explainable, are compatible, and that have the highest frequency of occurrence of expected reactions which match the test reactions.

Many of these goals and the kinds of knowledge that are used to attempt to achieve them are quite similar to those in the INTERNIST-1 method. For example, knowledge of the test reactions(data) expected to be present if a particular antibody (solution) is present are used to determine what test reactions the presence of an antibody would explain. INTERNIST-1 encoded similar knowledge of the manifestations(data) expected if an individual disease(solution) were present. These expectation are then used to achieve many similar goals to those in INTERNIST-1's method. For example, in a manner similar to INTERNIST-1 the match of expectations to the test reactions is combined in RED-2 with knowledge of the frequency of occurrence of these reactions to evaluate an individual antibody as a solution .

The similarities in the knowledge which maps expectation associated knowledge to evaluations of solutions is interesting because of the difference in encoding used. For example, degrees of plausibilities derived from matching individual manifestation expectations are dynamically combined using a fixed non-linear weighting scheme in INTERNIST-1. INTERNIST-1 therefore includes a set of operations and functions to utilize and manipulate these plausibilities. In RED-2, the knowledge mapping from the frequency of occurrence of matched expectations and plausibility of a solution is also static. However, in RED-2 this fixed weighting is represented as numerous instances of specific production rules. This evaluation knowledge is of the same kind and fills the equivalent evaluation role to knowledge that INTERNIST-I utilizes to evaluate diseases, although a cursory view of the form of the knowledge would lead one to conclude otherwise. The difference is more form then content, INTERNIST-1 computes a static mapping whereas RED-2 matches static structures to accomplish the mapping. On the one hand, INTERNIST-1 seems more flexible than RED-2 as it can combine the contribution of data patterns to solution evaluations dynamically. However, it is just a shorthand for the same static view of how knowledge related to various frequency measures of expectations can be mapped to solution evaluations.

This encoding difference is, we believe, related to the characteristics of the data in the two domains. In the RED-2 domain, the discriminable effects of the underlying processes giving rise to the test reactions are of a limited number Types of processes which could give rise to non-discriminable patterns can largely be ignored without effecting problem-solving accuracy. This makes it practical to pre-enumerate data patterns, frequency measures, and their desired effect on plausibility of a solution. In the INTERNIST-1 domain on the other hand, the possible combinations of datum are numerous given the more then 4,000 potential individual manifestations. It is therefore less practical for the INTERNIST-1 knowledge base to be created by pre-enumerating all the data patterns and their desired

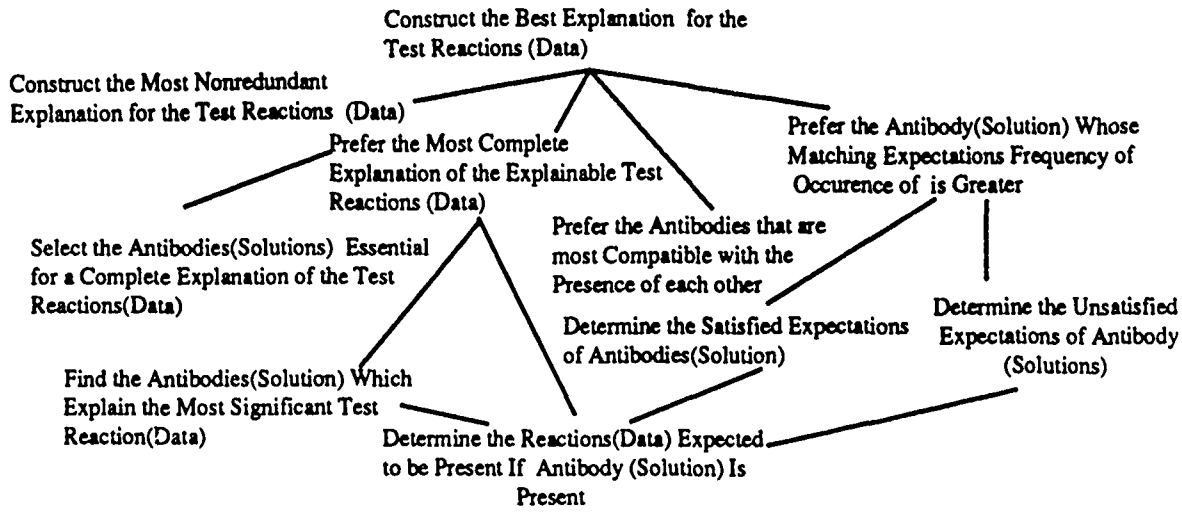


Figure 2. Problem-Solving Method: Goal/Subgoal Structure in Red

effect on plausibility. This route would translate into a large knowledge base without a clear means to exclude knowledge from application for a particular case.

RED-2, like INTERNIST-1, decomposes into possibly numerous instances the subgoal of constructing the best explanation into selecting the best individual antibodies. This decomposition is justified for this domain since the effect on the test reactions(data) of causal interactions between concurrently present antibodies is always additive, predictable, and easily computed. The search for a complete explanation is decomposed into possible numerous instances of explaining the most significant individual test reaction remaining to be explained. RED-2, like INTERNIST-1, encodes knowledge to determine the importance of explaining each test reaction value. The goal of non-redundant explanations is pursued by applying the knowledge of what each individual antibody can explain and knowledge of what combinations of antibodies can explain. As in INTERNIST-1, the data that are accounted for by the expectations of the current combination of accepted solution elements are considered explained and additional candidate solution elements are generated based on the remaining data left to explain.

This similarity in method is striking given the differences in the characteristics of the causal processes underlying the problems in the two domains. In the RED-2 domain, the causal processes giving rise to the test reactions are causally independent and give rise to largely discriminable patterns of data on the tests. This makes it justifiable to apply problem-solving methods which assume that overall goals can be decomposed as above. In the INTERNIST-1 domain, on the other hand, causal interactions are frequent and potentially numerous. It is therefore much less justifiable for the INTERNIST-1 designers to adopt such a decomposition strategy. There may be a robustness to the decomposition of such goals in the context of the multiple dimensions used for evaluation of potential candidate solutions that has not been previously appreciated.

Another apparent exceptions to the similarity in INTERNIST-1 and RED turns out to be more form then content. In RED-2 there is nothing that on the surface is comparable to INTERNIST-1's evoking strength that is used to evaluate a solution element. However, RED-2 searches for antibodies which are essential for explaining some reaction. Selecting essentials does not appear as an explicit subgoal in INTERNIST-1. It is interesting to see this notion does exist, implicitly, in INTERNIST-1. Let us assume we have a symptom that can be explained by only one disease. If the import of the symptom is below the threshold, it might never get explained and therefore the corresponding essential disease would not be included in the diagnosis. If the import is above the threshold however, it will be explained. As soon as the disease explaining this finding is ranked the highest, the system will consider it. Since there is only one disease to explain the given manifestation, the partitioning algorithm will isolate it (it has no competitor since it explains a datum no other disease can explain). INTERNIST-1 will either integrate the disease in the final diagnosis or will list it as a likely present next to the final diagnosis. Therefore, we can conclude that INTERNIST-1 has the

knowledge to qualify the essentials and that its problem-solving method will make sure that these essentials are included in the final diagnosis if at least some of the data they account for has an import greater than the threshold.

Conclusion

Our analysis of the knowledge content and problem-solving methods encoded in RED-2 and INTERNIST-1 has allowed us to characterize both systems in similar terms. Having couched both systems in the same vocabulary, we have been able to compare them more precisely. This comparison has yielded interesting insights on how similar each system's knowledge and goal structure are. We have argued that some goals and knowledge explicitly part of one system were implicit in the other. We can also clearly see how the form this knowledge is encoded in was influenced by characteristics intrinsic to the domains of blood banking and internal medicine. The impracticality of manually enumerating knowledge exhaustively in INTERNIST-1 motivated a finer goal decomposition and the application of different kinds of knowledge to different goals. For example, INTERNIST-1 explicitly includes the subgoals of preferring the most evoked disease and the disease with the least unsatisfied expectations. RED-2 does not explicitly include the these goals but only one equivalent goal of preferring the best hypothesis based on expectation frequencies. On the other hand, the form of representation in RED-2 does allows for the direct representation of the composition of general mappings of expectations with specific knowledge about expectations. Form does not follow content is the lesson we take away.

We believe such an analysis could be applied to other systems encoding abductive problem-solving with the goal of producing a taxonomy of abductive problems,problem-solving methods, and representations. Such a taxonomy would be useful in generalizing theoretical and experimental results and improving knowledge engineering practice for domain problems approached from the perspective of abductive problem-solving. For example, such a taxonomy would help the knowledge engineer in selecting the appropriate methods and knowledge representations when confronted with a domain problem.

Acknowledgements. This work has been supported by the National Library of Medicine under grant LM-04298 and by the National Heart Lung and Blood Institute under grant HL-38776. Computer facilities were enhanced through gifts from Xerox Corporation.

References

J. Josephson, B. Chandrasekaran, J. Smith, and M. Tanner. "A mechanism for forming composite explanatory hypotheses", in *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 17, no.3, 445-454, May/June 1987.

J. Smith, J. Svirbely, C. Evans, P. Strohm, J. Josephson and M. Tanner. "RED: A Red-Cell Antibody Identification Expert Module", *Journal of Medical Systems*, vol. 9, no. 3, 121-138, 1985.

R. A. Miller, M. D., H. E. Pople, Jr., Ph. D., and J. D. Myers, M. D. "INTERNIST-1, An Experimental Computer-Based Diagnostic Consultant for General Internal Medicine", *New England Journal of Medicine*, vol. 307, 468-476;1982.

H. E. Pople Jr, J. D. Myers, R. A. Miller. "DIALOG: A model of diagnostic logic for internal medicine", in *Proceedings of the Fourth International Joint Conference in Artificial Intelligence*, 1975.

H. E. Pople Jr. "The formation of composite hypotheses in diagnostic problem solving. An exercise in synthetic reasoning", in *Proceedings of the Fifth International Joint Conference in Artificial Intelligence*, 1030-1037, 1977.

A. Newell. "The Knowledge Level", *AI Magazine*, 1-19, Summer 1981.

Clancey, William J. "Heuristic Classification", Artificial Intelligence, vol. 27, 289-350;1985.

Explaining Unexpected Financial Results

Walter Hamscher

Price Waterhouse Technology Centre
68 Willow Road
Menlo Park, CA 94025
hamscher@pw.com or wch@ai.ai.mit.edu

1 Motivation

We wish to automatically construct plausible, parsimonious, and adequate explanations for unexpected financial results. We also wish to generate leading questions to help disambiguate among alternative explanations.

Figures 1–3 show a *very* simple example, abbreviated from that in [Kosy and Wise, 1984]. Figure 1 shows five equations relating eight financial and operational variables. Of these eight, we take the variables Unit Cost, Unit Price, and Volume to be exogenous.

Figure 1: Sample Financial Relations

Gross Margin	=	Sales − Production Cost
Sales	=	Volume × Unit Price
Variable Cost	=	Volume × Unit Cost
Production Cost	=	Variable Cost + Indirect Cost
Indirect Cost	=	Variable Cost × 15%

Suppose we observe that Gross Margin has decreased since last period. Some plausible explanations are shown in Figure 2. All three of the explanations propose disturbances in the exogenous variables Unit Price, Unit Cost, and Volume. All three are parsimonious: one disturbance to each exogenous variable is hypothesized in each, they mention only the variables through which the disturbance propagates, and they do not mention any irrelevant variables. The first two are adequate, in the sense that the hypothesized disturbance entails the result. The third (*) is inadequate because it does not entail what we wanted explained: if both Sales and Production Cost increased, then Gross Margin might have either increased, decreased, or stayed the same.

Figure 2: Potential Explanations

Unit Price decreased
→ Sales decreased
→ Gross Margin decreased

Unit Cost increased
→ Variable Cost increased
→ Production Cost increased
→ Gross Margin decreased

*Volume increased
→ Sales increased
→ Variable Cost increased
→ Indirect Cost increased
→ Production Cost increased

Because there are multiple explanations, we need to know what to ask next in order to disambiguate among them. Figure 3 shows four relevant queries. All four

Figure 3: Disambiguating Queries

Did Sales increase, decrease, or remain steady?
*Did Production Cost ...?
**Did Variable Cost ...?
**Did Indirect Cost ...?

queries are equally discriminating, since we could use the query result to rule out two of the explanations in Figure 2. However, we find the last two (**) to be undesirable, because neither quantity appears directly in the computation of Gross Margin, the anomaly that we are trying to explain. Also, the second query (*), like the last two, are undesirable queries because financial reports seldom report these quantities directly.

Protocol analysis has shown that this abductive style of reasoning manifests itself in tasks such as financial assessment [Bouwman, 1983], going concern evaluation [Selfridge *et al.*, 1986], and auditing [Dhar *et al.*, 1988, Peters, 1989]. Informal study indicates that it also plays a role in variety of other settings, such as tax planning, in which much of the effort is devoted to analyzing differences between prior cases and the case at hand. Facilities for abductive reasoning will play a key role in the Business Understander, the embodiment of a vision of a next generation knowledge-based computerized facility for supporting the understanding of client businesses by Price Waterhouse practitioners [Hamscher *et al.*, 1989].

The current example shows an essentially financial

model, in which all of the parameters are quantitative and most of them refer to amounts of money. Explanations referring only to dollar quantities may be adequate in a technical sense, but often unsatisfying for the tasks we envision; in the example above, a more satisfying explanation would refer to (say) increased competition to account for decreasing Unit Price. To generate such explanations requires operational models, in which the parameters refer to aspects of business such as the quality of the products, the lead time for new products, the brand loyalty of customers, and so forth [Hamscher, 1989]. We believe that the essential features of the abductive reasoning will remain unchanged in spite of the resulting shift in the character of the model.

The next two sections elaborate on the construction of explanations and discriminating queries in this domain. First, we present CROSBY, a reimplementation of the diagnosis engine SHERLOCK [de Kleer and Williams, 1989]. Next, we discuss approaches to the challenges encountered. These difficulties include the traditional issues arising from the interaction of a domain model with our abductive reasoning engine, as well as some nontraditional issues encountered arising from the difficulty of modeling business operations.

2 Implementation

The current implementation of CROSBY follows the traditional architecture of a model-based diagnosis program [Davis and Hamscher, 1988]:

Prediction There is a domain model that supports predictions about the behavior of the system under study; that is, given some facts about its behavior, the model predicts what behavior will be observed subsequently. In CROSBY this is based on local propagation of constraints [Sussman and Steele, 1980] over the domain of signs of each quantity and its first derivative with respect to time, as in many qualitative reasoning systems [Bobrow, 1985, Williams, 1988]. For example, if Unit Cost increases (denoted $[\partial \text{ UnitCost}] = [+]$) while Volume is constant ($[\partial \text{ Volume}] = [0]$) and both are positive ($[\text{UnitCost}] = [\text{Volume}] = [+]$) then Variable Cost will increase ($[\partial \text{ VariableCost}] = [+]$).

Hypothesis Space Definition The space of potential explanations is defined by the cross product of values that can be taken on by selected key variables. In other domains these key variables may be boolean-valued and refer to diseases [Reggia *et al.*, 1983], states of components in designed artifacts [de Kleer and Williams, 1987, Reiter, 1987], or they may be multiple-valued and refer to behavioral modes of components [de Kleer and Williams, 1989, Hamscher, 1990]. In CROSBY the key variables are simply the signs of the exogenous parameters of the financial model and their first derivatives. A prior probability is estimated for each variable assignment, with independence assumed among the parameters, and the resulting distribution summing to unity. For example, the parameter $[\partial \text{ Unit Cost}]$ may take on the values [0], [+], or [−] with prior probabilities .80, .15, and .05 respectively.

Conflict Detection Some combinations of key variable assignments are inconsistent and need to be ruled out. In CROSBY, as in SHERLOCK, each value in the domain of each key variable is associated with an ATMS assumption [de Kleer, 1986]. Each new deduction is given a label corresponding to the sets of minimal sets of assumptions needed to deduce that value. As pointed out in [Levesque, 1989], the construction of a label for a proposition constitutes an abductive inference for that proposition with respect to implicit beliefs (that is, the underlying assumptions). Hence by propagating labels through a network of the Horn clauses resulting from inferences made by the constraint propagator, explanations are constructed for every explicit belief (that is, the assignments of values to variables). When values deduced using different assumptions disagree, the set of underlying assumptions is declared to be in conflict. All supersets of that conflict set are inconsistent.

Interpretation Construction An interpretation is a consistent set of assumptions that is maximal in the sense that no assumption can be added to it without making it inconsistent. SHERLOCK performs heuristic best-first search through the space of interpretations. The evaluation function for each interpretation is the upper bound of its prior probability, assuming independence among its key variables, conditioned on any observations made so far, normalized with respect to all interpretations, and with an evaluation of 0 assigned to any interpretation discovered to be inconsistent.

For each set of assumptions ("environment"), there is a network of Horn clauses whose conclusions are supported in that environment. These are called the "active" clauses of that environment. Each network of active clauses supports some observations that the user may wish to see an explanation for given a certain environment. Each fact (such as $[\partial \text{ Sales}] = [-]$) may be supported by several active clauses, and for clarity it is best to select just one to display to the user. Three local criteria are used to make this choice among the supporting clauses of the fact: (i) The clause chosen must be active in the interpretation that the user is examining (ii) the clause must be that which is active in the environment of smallest cardinality, and (iii) the clause must have the fewest antecedent facts. Recursive application of these local criteria within an interpretation yields a directed graph of active clauses that forms an explanation structure. There is no guarantee of global economy in terms of facts or clauses, but the results are easily comprehensible in practice. Examples were shown in Figure 2. CROSBY displays the directed graph omitting facts in which variables are asserted to be unchanged, such as $[\partial \text{ Volume}] = [0]$.

Query Generation Some set of model parameters are declared to be observable. The model will have deduced one or more values for each such parameter, depending on different sets of assumptions. Since a probability is associated with each interpretation, and ideally each interpretation assigns a value to each observable parameter, a conditional probability for each value given each interpretation can be computed, and Bayesian diagnosis can be used to select the observation with the most information [de Kleer and Williams, 1987]. Let $p_{ik} = p(V_i = O_{ik})$, the probability that the observation of V_i has outcome O_{ik} conditioned on all observations made so far. Then the best observation on average is that with the minimum Shannon entropy $E_i = \sum_k p_{ik} \log p_{ik}$.

In this domain, variables corresponding to financial statement items (Sales, Gross Margin) are easily observable, certain operational variables (Volume) can be observed with varying degrees of difficulty, and the remainder are virtually impossible to observe directly because normal accounting systems to not record them as such. An extension of the entropy-based scheme estimates the desirability of observations having varying costs. The program selects the observation with minimum expected total cost $T_i = c_i + C_i(E_i - D)$, where:

- c_i is the cost of observing V_i,
- D is the entropy of the current set of interpretations I, that is, $D = \sum_j p(I_j) \log p(I_j)$,
- $(E_i - D)$ estimates the residual uncertainty after observing V_i,
- C_i is the average cost of observing the remaining unobserved variables, $C_i = \sum_{k \neq i} c_k$.

As intended, CROSBY generally selects observations with minimum c_i, except when there is a more expensive observation that would discriminate more strongly among competing interpretations.

3 Discussion

Test cases run with CROSBY highlight some issues both familiar to those working in the area of automated diagnosis, as well as some that are unique to the domain.

Ambiguity and Incompleteness Qualitative reasoning has the advantage of allowing a discrete range of values for key parameters and hence a finite space of explanations. However, it sometimes fails to detect inconsistent interpretations. For example, the interpretation corresponding to the *'d explanation in Figure 2 is inconsistent with [∂ Gross Margin] = [–] if [Volume] = [+], but the propagator fails to detect this. There are numerous known approaches to this problem involving quantity spaces, integrated numeric and qualitative approaches, and so forth. In the short term CROSBY uses the straightforward approach of employing multiple formulations of the same underlying model. For example, if the additional constraints in Figure 4 are added, the inconsistency would be discovered upon observing [∂ Unit Margin].

Figure 4: Additional Financial Relations

Gross Margin	=	Volume × Unit Margin
Unit Margin	=	Unit Price – Unit Cost

Combinatorics It is worrisome that the number of interpretations is exponential in the number of key variables, since the SHERLOCK procedure can in principle explore them all. However, since they are explored in the order of the likeliest interpretations first, we believe that the performance will be acceptable in practice, and this is supported by some empirical evidence. The Symbolics 3645 implementation of SHERLOCK by its original authors already diagnoses combinational digital circuits with up to a hundred gates in a matter of seconds [B. C. Williams, personal communication]. CROSBY is also implemented on a Symbolics 3645, but its underlying ATMS is not as fast. Its largest example to date involves 15 variables, 6 of them exogenous, and completes in 82 seconds. This 82 seconds does not include the I/O time to solicit four observations from the user.

Incoherent Lines of Questioning In the applications that we envision, the role of the abduction engine is to help the human user formulate plausible theories and subsequent lines of investigation [Hamscher *et al.*, 1989]. Bayesian diagnosis is notorious for giving the user an uncomfortable sense of "jumping around" between different hypotheses [Szolovits and Pauker, 1978, Pople, 1982]. We plan to explore several different approaches to this latter issue:

Use Guided Probe Perhaps Bayesian diagnosis is simply irrelevant when the goal is comprehensibility rather than optimality. For example, in the domain of digital circuit troubleshooting, the "guided probe" algorithm involves a methodical step-by-step upstream tracing of causal paths; while the resulting sequences are suboptimal, it requires that little state be retained and as a result is easy to follow.

Use More Layers Perhaps the sense of "jumping around" is simply an artifact of having constructed explanations with too many steps, which in turn is an artifact of having descended to too low a level of detail. Suppose that the abductive engine were extended to perform hierarchic diagnosis, and consider the earlier example: if the model had several layers, each one involving only a small increase in the number of variables, then as long as the abductive engine stayed at a single level there would be less jumping around — simply because there would be fewer variables of interest at any given level, hence fewer places to jump to. Thus, perhaps the problem is not Bayesian diagnosis, but rather that the models used are not richly enough layered.

Dynamically Bias Observation Costs Perhaps Bayesian diagnosis could yield the same sense of controlled focus as the guided probe algorithm. The estimated cost of each new observation could be dynamically adjusted to prefer those observations that are more closely related to the most recent observations made. For example, having observed the variable [∂ Gross Margin], the next observation would be biased toward [Gross Margin], and toward [∂ Production Cost] or [∂ Sales] because these variables appear in a relation with Gross Margin (Figure 1).

The current implementation of CROSBY strikes a balance between (myopic) guided probing and (global) Bayesian diagnosis. CROSBY defines the cost of observing V_i to be $c_i' = (1-\alpha)c_i + \alpha d_i$, where d_i is the number of relations intervening between V_i and the most recently observed variable, and α is a "locality bias" parameter ranging between 0 (global) and 1 (myopic). For computational simplicity, the observation is selected to minimize $T_i' = c_i' + C_i(E_i - D)$. However, the models constructed up to this time have not been large enough to provide a useful test of this scheme.

Finance is Not Enough As noted earlier, an explanation of declining gross margins that simply refers to prices, costs, and volumes is fundamentally unsatisfying. What explains *them*? Most current Artificial Intelligence work in model-based reasoning is grounded in classical physics. Reasoning tasks involving business operations are not grounded in physics, but rather in the disciplines of economics, accounting, marketing, human resource management, and so forth. These disciplines are rich in quantitative methods, and appeal to similar types of assumptions (continuity, linearity, closed worlds, and so forth), but do not approach the breadth and predictive power of classical physics.

A short term technical issue that thus arises is that discrepancies between predictions and observations (or between alternative predictions) should not necessarily have the status of conflicts (in the ATMS sense). For example, the factor of 15% mentioned in Figure 1 is merely an estimate of the relationship between Variable and Indirect costs, and variations are to be expected. In general, some discrepancies are much more significant than others, and the abductive engine should take this into account in focusing its efforts.

The long term technical issue arises from the need for comprehensive behavioral models of businesses, and the fact that many predictive theories in economics are currently based on regression analysis. Some would claim that the proposed use of regression based schemes is misguided, according to the following argument: building such a model requires one to have some kind of causal theory before one can write the equations down. An extremely important reason for choosing carefully the structure of the equations is to reduce the effects of interdependencies among factors. Hence causal assertions like "the number of potential customers and their ability to purchase was considered an important external determinant of sales" [Elliott and Uphoff, 1972] justify the inclusion of one or another factors in a given equation, in this case the inclusion of GNP as a contributing factor to the sales of a given firm. Unfortunately, achieving a completely independent set of parameters is impossible in general and in any specific case leaves much room for debate: for example, the equation for prices in [Elliott and Uphoff, 1972] is a function of capital expenditures, rather than production costs, even though the latter seem to be more directly related to the way firms set their prices. Assuming that the model can be justified, subsequent regression will tell one what the parameters are, and tell one how well the historical data fit them. But what does one do if the fit is poor — conversely, what does it really mean if the fit is good? The causal story is no longer explicit in the equations, yet that was the background against which one must do all debugging and interpretation. In other words, if one uses the result of a regression (or related techniques such as smoothing) to make a prediction, how should one interpret data that either agrees or disagrees with the prediction? These difficulties argue against the use of regression-based extrapolation, but it appears that the field of economics currently offers few alternatives.

4 Conclusion

We wish to automatically construct plausible, parsimonious, and adequate explanations for unexpected financial results. We have constructed an initial prototype abductive engine based on a model-based diagnosis engine and are exploring several issues in modeling, diagnosis, and explanation.

Acknowledgements

The author acknowledges discussions with Richard Fikes, Beau Sheil, and Alan Timmins. Luisa Claeys tracked down the name of Sherlock Holmes' banker: Crosby.

References

[Bobrow, 1985] D. Bobrow, editor. *Qualitative Reasoning about Physical Systems.* MIT Press, Cambridge, MA, 1985.

[Bouwman, 1983] M. J. Bouwman. Human Diagnostic Reasoning by Computer: An Illustration from Financial Analysis. *Management Science*, 29(6):653–672, June 1983.

[Davis and Hamscher, 1988] R. Davis and W. C. Hamscher. Model-based Reasoning: Troubleshooting. In H. E. Shrobe, editor, *Exploring Artificial Intelligence: Survey Talks from the National Conferences on Artificial Intelligence*, pages 297–346. Morgan Kaufmann, San Mateo, CA, 1988.

[de Kleer and Williams, 1987] J. de Kleer and B. C. Williams. Diagnosing Multiple Faults. *Artificial Intelligence*, 32(1):97–130, April 1987.

[de Kleer and Williams, 1989] J. de Kleer and B. C. Williams. Diagnosis with Behavioral Modes. In *Proc. 11th Int. Joint Conf. on Artificial Intelligence*, pages 1324–1330, Detroit, MI, 1989.

[de Kleer, 1986] J. de Kleer. An Assumption-Based TMS. *Artificial Intelligence*, 28(2):127–162, 1986.

[Dhar *et al.*, 1988] V. Dhar, B. Lewis, and J. Peters. A Knowledge-Based Model of Audit Risk. *AI Magazine*, 9(3):57–63, Fall 1988.

[Elliott and Uphoff, 1972] J. W. Elliott and H. L. Uphoff. Predicting the Near Term Profit and Loss Statement with an Econometric Model: A Feasibility Study. *Journal of Accounting Research*, pages 259–274, Autumn 1972.

[Hamscher *et al.*, 1989] W. C. Hamscher, R. E. Fikes, and A. Timmins. The Business Understander. Price Waterhouse Technology Centre Working Paper, October 1989.

[Hamscher, 1989] W. C. Hamscher. Business as a Domain for Model-based Reasoning: A Proposal to Use Explicit Representation of the Structure and Behavior of Business Entities to Assess Audit Risk. Proc. IJCAI Workshop on Model-based Reasoning, August 1989. Available from the author, Price Waterhouse Technology Centre, 68 Willow Road, Menlo Park, CA 94025.

[Hamscher, 1990] W. C. Hamscher. Diagnosing Devices with Hierarchic Structure and Known Component Failure Modes. In *Proc. 6th IEEE Conf. on A.I. Applications*, Santa Barbara, CA, March 1990.

[Kosy and Wise, 1984] D. W. Kosy and B. P. Wise. Self-Explanatory Financial Models. In *Proc. 4th National Conf. on Artificial Intelligence*, pages 176–181, Austin, TX, August 1984.

[Levesque, 1989] H. J. Levesque. A Knowledge-level Account of Abduction. In *Proc. 11th Int. Joint Conf. on Artificial Intelligence*, pages 1061–1067, Detroit, MI, August 1989.

[Peters, 1989] J. M. Peters. *A Knowledge Based Model of Inherent Audit Risk Assessment.* PhD thesis, Katz Graduate School of Business, University of Pittsburgh, 1989. Available from the author, Department of Accounting, College of Business Administration, University of Oregon, Eugene, OR 97403-1208.

[Pople, 1982] H. E. Pople. Heuristic Methods for Imposing Structure on Ill-structured Problems: The Structuring of Medical Diagnostics. In P. Szolovits, editor, *Artificial Intelligence in Medicine*, pages 119–190. Westview Press, Boulder, CO, 1982.

[Reggia *et al.*, 1983] J. A. Reggia, D. S. Nau, and P. Wang. Diagnostic Expert Systems Based on a Set Covering Model. *Int. Journal of Man-Machine Studies*, 19(5):437–460, November 1983.

[Reiter, 1987] R. Reiter. A Theory of Diagnosis from First Principles. *Artificial Intelligence*, 32(1):57–96, April 1987.

[Selfridge *et al.*, 1986] M. Selfridge, S. F. Biggs, and G. R. Krupka. GCX: A Cognitive Expert System for Making Going-Concern Judgements. Technical Report 86-20, University of Connecticut Department of Computer Science and Engineering and Computer Applications and Research Center, Storrs, CT 06268, December 1986.

[Sussman and Steele, 1980] G. J. Sussman and G. L. Steele. Constraints: A Language for Expressing Almost-hierarchical Descriptions. *Artificial Intelligence*, 14(1):1–40, January 1980.

[Szolovits and Pauker, 1978] P. Szolovits and S. G. Pauker. Categorical and Probabilistic Reasoning in Medical Diagnosis. *Artificial Intelligence*, 11:115–144, 1978.

[Williams, 1988] B. C. Williams. MINIMA: A Symbolic Approach to Qualitative Algebraic Reasoning. In *Proc. 7th National Conf. on Artificial Intelligence*, pages 264–269, Minneapolis, MN, August 1988.

Abductive Solutions to Temporal Projection Problems

Murray Shanahan

Imperial College
Department of Computing,
180 Queen's Gate,
London SW7 2BZ.
England.

Introduction

After Hanks and McDermott demonstrated that formalising default persistence to overcome the frame problem was less straightforward than it at first seemed [Hanks and McDermott, 1987], several authors developed more robust formalisations, such as Lifschitz [1987] who uses circumscription, Shoham [1988] who uses model preference, and Evans [1989] who uses negation-as-failure. These formalisations can describe the so-called Yale shooting scenario which Hanks and McDermott introduced, and they yield the intended predictions. However, these formalisations of default persistence need to be modified to cope with Kautz's stolen car problem [Kautz, 1986], and the bloodless variation of the Yale shooting scenario. A few authors have proposed solutions to these problems, such as [Morgenstern and Stein, 1988], [Lifschitz and Rabinov, 1989] and [Shanahan, 1989a].

In [Shanahan, 1989a], a solution to Kautz's stolen car problem is suggested which uses abduction. However, the paper concentrates on the issue of explanation in temporal reasoning, and the proposed solution is not discussed in depth. Also, the proposed solution uses negation-as-failure to achieve default persistence. This apparently restricts the solution to extended Horn clause representations of change, such as the shortened version of Kowalski and Sergot's Event Calculus used in the paper. But the principle behind the solution is equally applicable to full first-order predicate calculus representations. This paper seeks to present the abductive solution to both Kautz's stolen car problem and the bloodless Yale shooting problem, using a version of the Event Calculus which employs circumscription rather than negation-as-failure.

1. A Circumscriptive Event Calculus

The different versions of the Event Calculus presented in, for example, [Kowalski and Sergot, 1986], [Kowalski, 1986] and [Shanahan, 1989a], all use negation-as-failure to achieve default persistence. This paper uses a variation of the Event Calculus axioms, and employs circumscription to achieve default persistence. The axioms are as follows. All variables are universally quantified unless otherwise shown.

$$holds\text{-}at(P,T2) \leftarrow \quad \text{(E.1)}$$
$$\exists E,T1\ [happens(E) \wedge success(E) \wedge time(E,T1) \wedge T1 < T2 \wedge initiates(E,P) \wedge \neg\, clipped(T1,P,T2)]$$

$$clipped(T1,P,T3) \equiv \quad \text{(E.2)}$$
$$\exists E,T2\ [happens(E) \wedge success(E) \wedge time(E,T2) \wedge terminates(E,P) \wedge T1 \leq T2 \wedge T2 < T3]$$

$$success(E) \equiv \quad \text{(E.3)}$$
$$time(E,T) \rightarrow \forall P[precond(E,P) \rightarrow holds\text{-}at(P,T)]$$

$$time(E,T1) \wedge time(E,T2) \rightarrow T1=T2 \quad \text{(E.4)}$$

$$act(E,A1) \wedge act(E,A2) \rightarrow A1=A2 \quad \text{(E.5)}$$

These axioms are very similar to those used in [Shanahan, 1989a]. The basic ontology includes events, time points and properties. The predicate *happens(E)* represents that event E occurs, *time(E,T)* represents that E occurs at time T, *act(E,A)* represents that event E is of type A and *holds-at(P,T)* represents that property P holds at time T. The predicate *clipped(T1,P,T2)* represents that property P is terminated at some time between times $T1$ and $T2$. The domain is represented by the predicates *initiates* and *terminates*. Respectively, *initiates(E,P)* and *terminates(E,P)* represent that the property P is initiated by the event E and terminated by the event E. The predicate *precond(E,P)*, which is taken from [Lifschitz, 1987], represents that P must hold at the time of event E for it to have any effect.

To achieve default persistence, whereby a property, once initiated, persists by default until an event occurs which terminates it, we circumscribe the theory $E \cup D \cup H$, where E is the theory comprising (E.1) to (E.5) (and including an implicit equality and inequality theory), D is a set of axioms defining *initiates*, *terminates* and *precond* and H is a set of axioms describing a history of events in terms of *happens*, *act* and *time*. The circumscription policy minimises

happens, initiates, terminates and *precond*, and also *holds-at* with a lower priority.

Circum(E ∪ D ∪ H ; holds-at < happens, initiates, terminates, precond)

As in [Lifschitz, 1987], the introduction of the *precond* predicate allows the preconditions of axioms to be minimised independently of anything which varies over time. Lifschitz also introduced a *causes* predicate for the similar reasons, and the *initiates* and *terminates* predicates of the Event Calculus serve the same purpose. An alternative formulation results if *holds-at* is not minimised, and Axiom (E.1) is written as an equivalence rather than an implication. But this form of Axiom (E.1) excludes the possibility of discovering other ways in which a property can hold. In [Shanahan, 1989b], for example, an addditional *holds-at* axiom is introduced to cope with continuous change. But I will return to this alternative formulation later, because it also yields an apparently straightforward solution to the bloodless Yale shooting problem and Kautz's stolen car problem.

There is a potentially serious problem with this circumscriptive formulation of the Event Calculus which I don't propose to tackle in this paper, but which needs to be pointed out. As well as extra *holds-at* axioms to describe continuous change, most domains require the addition of extra *holds-at* axioms defining non-primitive properties, that is, properties which are not initiated and terminated themselves, but which are derived from those which are. For example, in the Blocks World, we might define the property *clear(X)* to hold if nothing is on top of the block *X*, where the property *on(Y,X)*, representing that block *Y* is on top of *X*, is a primitive property initiated and terminated by events. Unfortunately, whilst the formulation given will work well with many axioms defining non-primitive properties, some such axioms will give rise to unexpected models. This difficulty also arose for Lifschitz [1987], who proposes a solution which may carry over in some form to the formulation given here.

2. The Yale Shooting Problem

The simple domain of the Yale shooting problem comprises only three types of event — *load*, *wait* and *shoot* — and three properties — *alive*, *dead* and *loaded*. A fourth type of event — *birth* — is introduced to comply with the basic intuition behind the Event Calculus that all properties which hold must have an explanation in terms of events. These events and properties are axiomatised as follows. Note that there are no axioms for the *wait* event, since it has no effect.

initiates(E,loaded) ← act(E,load)	(D1.1)
initiates(E,dead) ← act(E,shoot)	(D1.2)
initiates(E,alive) ← act(E,birth)	(D1.3)
terminates(E,alive) ← act(E,shoot)	(D1.4)
precond(E,loaded) ← act(E,shoot)	(D1.5)

Now, the Yale shooting scenario can be represented by the following history, comprising four events — birth then load then wait then shoot.

happens(e0)	(H1.1)	*happens(e1)*	(H1.4)
time(e0,t0)	(H1.2)	*time(e1,t1)*	(H1.5)
act(e0,birth)	(H1.3)	*act(e1,load)*	(H1.6)
t1 > t0	(H1.7)		
happens(e2)	(H1.8)	*happens(e3)*	(H1.12)
time(e2,t2)	(H1.9)	*time(e3,t3)*	(H1.13)
act(e2,wait)	(H1.10)	*act(e3,shoot)*	(H1.14)
t2 > t1	(H1.11)	*t3 > t2*	(H1.15)
t4 > t3	(H1.16)		

In all models of the circumscription of $E \cup D1 \cup H1$, minimising *happens*, *initiates*, *terminates* and *precond*, and then minimising *holds-at* with a lower priority, we have *holds-at(loaded,t3)* and therefore *holds-at(dead,t4)* and *–holds-at(alive,t4)*. Anomalous models, in which the gun is unloaded by the *wait* event or by some other event, do not arise because they are less minimal in either *happens* or *terminates* than models in which the *loaded* property persists.

3. The Bloodless Yale Shooting Problem

In the bloodless variation of the Yale shooting problem, we have the same history of events — namely load then wait then shoot — but we are also told that *alive* holds afterwards. Somehow, the addition of this fact must block the inference that *dead* holds as a result of the shoot event. There are two ways to view this problem. We might consider that the addition of this new fact simply results in a different prediction problem, which understandably produces different predictions. Since the predictions rely on default persistence, which is non-monotonic, there is nothing strange about the fact that the addition of a new fact results in the retraction of a previous prediction. Alternatively, we might regard the new fact as requiring an explanation. On this latter view, it is not sufficient simply to derive new predictions from the new fact. Rather, we have to seek possible explanations for the new fact, and only derive new predictions from these explanations.

The first approach might be realised by expressing Axiom (E.1) as an equivalence rather than an implication. Then, the addition of *holds-at(alive,t4)* to *H1* implies that nothing happened before *t4* to clip alive. But since the shoot event *e3* happened before *t4*, and since shooting terminates *alive*, either the shooting was unsuccessful, or a second birth event took place which isn't mentioned in the history, or a completely new type of event took place which initiates alive, say a resurrection event. The first possibility — that the shooting was unsuccessful — implies that either the wait event unloaded the gun, or that some other event happened to unload the gun which wasn't mentioned in the

history. In short, as well as requiring an ammendment to the axioms which could turn out to be false, adding *holds-at(alive,t4)* to *H1* yields only a very weak disjunction. Incidentally, this disjunction does not exclude the possibility that *dead* also holds at *t4*, but this could be rectified by adding a suitable axiom to *D1*.

The second approach regards the problem as having an explanation component as well as a prediction component. It is not appropriate simply to add *holds-at(alive,t4)* to *H1* and then to derive new predictions. Rather, *holds-at(alive,t4)* has to be explained. The set of possible explanations of this fact can then be added to *H1*, generating a set of ammended histories, and new predictions can be made with these new histories. From a logical point of view, the problem involves an abductive component for the explanation and a deductive component for the prediction. The abductive component is to find an explanation Δ which is consistent with $E \cup D1 \cup H1$ such that *holds-at(alive,t4)* is in all models of the circumscription

Circum(E ∪ D1 ∪ H1 ∪ Δ ; holds-at < happens, initiates, terminates, precond)

and the deductive component is to make predictions by finding those sentences which are in all models of the such circumscriptions. Of course, there may be many Δ's to explain any given fact. So the definition of an acceptable explanation is further refined. A set of predicates is distinguished as the *abducibles*. In general, given a domain theory D and a history H, an acceptable explanation Δ for a fact G is a set of atomic sentences involving only abducible predicates, such that G is in all models of the circumscription

Circum(E ∪ D ∪ H ∪ Δ ; holds-at < happens, initiates, terminates, precond)

and there is no Δ^* comprising atomic sentences involving only abducible predicates such that $\Delta \supset \Delta^*$ and G is also in all models of the circumscription

Circum(E ∪ D ∪ H ∪ Δ ; holds-at < happens, initiates, terminates, precond)*

There may still be many acceptable Δ's to explain a given fact. The set of sentences which is the intersection of all such Δ's is called the set of *defeasibly necessary conditions* for the fact. Each individual Δ is a set of *defeasibly sufficient conditions* for the fact. Finally, a further preference relation << can be defined on Δ's, which captures the idea of a good explanation. For example, Δ's may be preferred which introduce fewer new events — $\Delta_1 << \Delta_2$ iff the set of all *happens* sentences in Δ_1 is a proper subset of the set of all *happens* sentences in Δ_2. A *preferred* explanation is any acceptable explanation Δ such that there is no acceptable explanation Δ^* where $\Delta^* << \Delta$. Note that an explanation may introduce constants which do not appear in $E \cup D \cup H$, to name new events for example. Explanations which differ only in the symbols they use for new constants are considered the same.

This formal apparatus allows us to tackle a variety of temporal explanation problems. For the bloodless Yale shooting problem, we want explanations for *holds-at(alive,t4)*. But the question arises, which predicates do we make abducible. For many explanation problems, it is appropriate to make the temporal ordering predicates abducible along with the predicates *happens*, *time* and *act*. That is, the only abducible predicates are those which can feature in a history of events. If these are the only abducible predicates, then given the domain *D1* which does not include any types of event which can unload the gun, the only preferred explanation for *holds-at(alive,t4)* involves a reincarnation: *{happens(e), time (e,t), act(e,birth), t < t4, t3 < t}*. Making domain predicates abducible gives a different result. In particular, if we make *terminates* abducible, the only preferred explanation is *{terminates(e2,loaded)}*, that is the wait event unloads the gun.

4. The Stolen Car Problem

Kautz [1986] posed the following problem. Suppose that I park my car in the morning and go to work. At lunch time, without going to look at my car, I might reasonably apply default persistence and infer that the car is still where I left it. However, when I return to the car park in the evening I find that it has been stolen. My previous conclusion that the car was still there at lunch time is clearly now open to question. The car may have been stolen at any time after I parked it and before I observed that it was gone, so I cannot say anything about its whereabouts at lunch time. Any formalisation of default persistence should deal satisfactorily with this kind of scenario.

The domain can be trivially formalised as follows, using two types of event — *park* and *steal* — and the single property *in-car-park*.

initiates(E,in-car-park) if act(E,park) (D2.1)
terminates(E,in-car-park) if act(E,steal) (D2.2)

There is only one event, that of parking the car, which we can represent as follows.

happens(e0) (H2.1)
act(e0,park) (H2.2)
time(e0,morning) (H2.3)

We also have

morning < lunch-time (H2.4)
lunch-time < evening (H2.5)

With these axioms alone, in all models of the circumscription of $E \cup D2 \cup H2$, minimising according to the policy in Section 1, default persistence gives us *holds-at(in-car-park,lunch-time)* and *holds-at(in-car-park,evening)*. However, if we now seek an explanation for ¬*holds-at(in-*

car-park,evening), we find that the only preferred explanation is Δ = *{happens(e), act(e,steal), time(e,t), morning* < *t*, *t* < *evening}*. Circumscribing $E \cup D2 \cup H2 \cup \Delta$ yields only models in which we have ¬*holds-at(in-car-park,evening)*, but because the relative ordering of *t* and *lunch-time* is not known, we have some models which include *holds-at(in-car-park,lunch-time)* and others which include *holds-at(in-car-park,lunch-time)*. In other words, we cannot conclude anything about the whereabouts of the car at lunch time.

Does this constitute a solution to Kautz's stolen car problem, or is it cheating? The acceptability of the abductive approach to such problems hinges on a view of knowledge assimilation which goes beyond the idea of simply adding new facts directly into the knowledge base. Suppose that we have a knowledge base in the form of a set of sentences *T*. Under a classical view of knowledge assimilation, new facts are added directly to *T*. With an abductive view of knowledge assimilation, new facts are added to the set of logical consequences *G* of the knowledge base, demanding the addition of a set of sentences Δ to *T* such that $T \cup \Delta \models G$. That is, each new fact must be explained. Using abduction with the Event Calculus, assimilating a new *holds-at* fact, such as the fact that my car is not in the car park in the evening, demands the addition of a whole set of *happens*, *time*, *act* and temporal ordering sentences, so that the new fact becomes a logical consequence of the knowledge base. With the stolen car problem, there is a unique preferred explanation, but this is not necessarily the case. Complications arise when there are many preferred explanations for a fact, but I will not address this problem here. One approach is to add the disjunction of all the most preferred explanations to the knowledge base.

5. Discussion

There are two important issues arising from the abductive approach to temporal projection which merit further discussion. The first concerns the relationship between approaches which mix circumscription (or some other form of default reasoning) for default persistence with abduction for explanation, and approaches which employ abduction both for default persistence and explanation. The second concerns the viability of an approach which uses only deduction plus circumscription (or some other form of default reasoning).

Some authors have recommended abduction as an approach to default reasoning in general [Poole, 1988], [Eshghi and Kowalski, 1989], [Kakas and Mancarella, 1989]. So it would seem to be possible to define an abductive framework which can deal with both default persistence and explanation. In addition to making *happens*, *time*, *act* and temporal ordering predicates abducible, we can define *persists* as ¬*clipped*, and make *persists* abducible [Eshghi, 1988]. It may turn out that this or a similar abductive approach to default persistence (such as [Goebel and Goodwin, 1987]) is equivalent to the circumscriptive approach, but this is a topic for further work. Even if abduction can be used for both default reasoning and explanation, it may still be important to keep these two uses of abduction conceptually distinct.

As discussed above, another approach to the temporal projection problems tackled in this paper is to use deduction with circumscription or some other form of default reasoning [Morgenstern and Stein, 1988], [Lifschitz and Rabinov, 1989]. This might be achieved by writing Axiom (E.1) as an equivalence instead of an implication. There are several objections to this approach, but none of them seem overwhelming. First, it might require writing equivalences which are, strictly speaking, false and should be written as implications. Second, it seems counter-intuitive to view explanation problems like the bloodless Yale shooting problem and the stolen car problem as deductive. But then again, any method which uses circumscription can't really be viewed as purely deductive anyway. Third, the large disjunctions which can result from this approach don't convey as much information as the set of defeasibly necessary conditions plus the set of sets of defeasibly sufficient conditions which are obtained with abduction. A thorough exploration of the relationship between these two approaches is another topic for further work.

Finally, it is worth adding a few words about implementation. Axioms (E.1) to (E.6) were derived from the axioms presented in [Shanahan, 1989a], which are all extended Horn clauses. It is quite straightforward to compile Axioms (E.1) to (E.6) back into extended Horn clauses, and we might expect this to be true of most extensions to the axioms. When they are expressed in extended Horn clause form, a simple extension to resolution will perform abduction, and a Prolog interpreter will suffice for deduction [Shanahan, 1989a].

Acknowledgements

Thanks to Kave Eshghi, Chris Evans, Tony Kakas, Bob Kowalski, Marek Sergot and Sury Sripada for many helpful discussions on abduction and temporal reasoning.

References

[Eshghi, 1988] K.Eshghi, Abductive Planning with Event Calculus, *Proceedings 5th International Conference on Logic Programming* (1988), p 562.

[Eshghi and Kowalski, 1989] K.Eshghi and R.A.Kowalski, Abduction Compared with Negation by Failure, *Proceedings 6th International Conference on Logic Programming* (1989), p 234.

[Evans, 1989] C.Evans, Negation-As-Failure As an Approach to the Hanks and McDermott Problem, *Proceedings 2nd International Symposium on Artificial Intelligence* (1989).

[Goebel and Goodwin, 1987] R.G.Goebel and S.D.Goodwin, Applying Theory Formation to the Planning Problem,

Proceedings of the 1987 Workshop on the Frame Problem, p 207.

[Hanks and McDermott, 1987] S.Hanks and D.McDermott, Nonmonotonic Logic and Temporal Projection, *Artificial Intelligence*, vol 33 (1987), p 379.

[Kakas and Mancarella, 1989] A.C.Kakas and P.Mancarella, Anomalous Models and Abduction, *Proceedings 2nd International Symposium on Artificial Intelligence* (1989).

[Kautz, 1986] H.Kautz, The Logic of Persistence, *Proceedings AAAI 86*, p 401.

[Kowalski, 1986] R.A.Kowalski, Database Updates in the Event Calculus, Imperial College Department of Computing Technical Report no. DOC 86/12 (1986).

[Kowalski and Sergot, 1986] R.A.Kowalski and M.J.Sergot, A Logic-Based Calculus of Events, *New Generation Computing*, vol 4 (1986), p 267.

[Lifschitz, 1987] V.Lifschitz, Formal Theories of Action, *Proceedings of the 1987 Workshop on the Frame Problem*, p 35.

[Lifschitz and Rabinov, 1989] V.Lifschitz and A.Rabinov, Miracles in Formal Theories of Action, *Artificial Intelligence*, vol 38 (1989), p 225.

[Morgenstern and Stein, 1988] L.Morgenstern and L.A.Stein, Why Things Go Wrong: A Formal Theory of Causal Reasoning, *Proceedings AAAI 88*, p 518.

[Poole, 1988] D.L.Poole, A Logical Framework for Default Reasoning, *Artificial Intelligence*, vol 36 (1988), p 27.

[Shanahan, 1989a] M.P.Shanahan, Prediction Is Deduction but Explanation Is Abduction, *Proceedings IJCAI 89*, p 1055.

[Shanahan, 1989b] M.P.Shanahan, Representing Continuous Change in the Event Calculus, Imperial College Department of Computing report, (1989).

[Shoham, 1988] Y.Shoham, Reasoning About Change: Time and Change from the Standpoint of Artificial Intelligence, MIT Press (1988).

Hypo-deductive reasoning for abduction, default reasoning and design

David Poole
Department of Computer Science,
University of British Columbia,
Vancouver, B.C., Canada V6T 1W5
poole@cs.ubc.ca

Abstract

Is a default more than just an assumed premise in a logical argument of what to predict? Is recognition just conjecturing causes that can be used as premises to imply the observations? Is design just choosing components that can provably do the job? An ongoing research programme by this author and others is to try to answer these questions. Rather than advocating complex sophisticated theories, we are trying to find out where simple solutions break down and only add complexity where it is needed. In this paper, I argue for this "minimalist" approach to AI, argue for a particular representational theory as an appropriate starting point, and then report on what we have found by such an endeavour. The starting point is a simple form of hypo-deductive reasoning where the user provides the forms of hypotheses they are prepared to accept as part of a logical argument.

1 Minimalist AI

As in any scientific endeavour we have to come up with theories. What should an AI theory look like? What should a theory of representation look like? This paper is an attempt to justify one representational theory.

A reasonable way to proceed is to do what could be called "minimalist AI". We only use tools that are demonstrably required, and only augment them when they are proven inadequate for the sort of reasoning we want to do. In this way many people have argued that any reasoning system should incorporate at least the first order predicate calculus.

One problem with this is that even a very weak logic (e.g., Horn clauses with function symbols) can represent any computable function. Therefore a theory that says all we need is logic is, at one level, vacuous. We already know, if AI is possible, we can represent intelligent reasoning with a Turing machine.

Just as a hammer is not just a piece of steel we can buy at a hardware store (we can *use* a rock as a hammer), an AI theory is more than the invention of a new logic, new formalism or even a new system. AI is about how to *use* tools to reason intelligently.

Ideally an AI theory must come up with a limited repertoire of tools and methodologies for using these tools. Under this analysis, research should consist of learning how to use this set of tools and only adding to it when it can be shown to be inadequate. We don't make advances by increasing the number or complexity of tools, but rather we make advances when we have fewer and simpler tools together with useful ways to use these tools.

2 Logic and Monotonicity

One of the primary arguments for using logic is that the notion of semantics is important (see discussions in [Levesque88]). The notion of logical consequence ($P \models c$) is usually taken to mean that c is true in all models of P. In other words there is no way that P can be true with c being false. If the conclusion c is false, then the premises P must be false. This primary tenet of logic can be summed up procedurally in the statement:

> *If you don't like the conclusion of a logical argument, don't criticise the logic, criticise the premises.*

Suppose we have a logical argument that Tweety flies based on Tweety being a bird, and "birds fly". If we subsequently learn that Tweety is an emu, then the conclusion (that Tweety flies) is wrong, but the logical proof is still valid. This problem of the "monotonicity" of logic has lead to many new "logics" to handle exceptions.

The logical argument is valid; we don't like the conclusion, so we should criticise the premises. What is the wrong premise? The wrong premise is "birds fly". This is a premise we don't want to use when the object under consideration is an emu. The idea that there are premises that we want to use some of the time, but not all of the time is the basis behind Theorist [Poole88a].

Logic tells us the consequences of our assumptions, it doesn't tell us where the assumptions come from. What should be the premises of a logical argument? The example above shows that they should be what we know (our "facts" — what we are not prepared to give up) together with hypotheses we are prepared to accept as part of an argument.

The Theorist conjecture is that we don't need anything more than this. We don't need new logics, new rules of inference, new semantics, we can do all reasoning in terms of theory formation using normal logic (the idea is independent of the logic, but we assume the first order predicate calculus as we can also argue that we need at least this if we want to represent indirectly described individuals, disjunction and explicit negation).

Where should our "hypotheses" come from? The answer to this is "I don't know". How should we go about answering this question? It seems as though the simplest idea is to allow the user to be able to provide the forms of assumptions they are prepared to accept as part of an explanation. By using such a system we can learn the principles behind such possible hypotheses. There seems to be no way, a priori, to say that something should or should not be a hypothesis; it is only by gaining experience that we will learn this. For a start we let the user provide the form of the hypotheses.

3 Theorist

Theorist [PGA87, Poole88a] is defined in terms of:

F a set of closed formulae, called the "facts"; these are regarded as true of the domain under consideration. As such, they are assumed to be consistent.

H a set of possible hypotheses, instances of which can be used as premises of a logical argument.

Definition: A **scenario** is $F \cup D$ where D is a set of ground instances of elements of H such that $F \cup D$ is consistent.

A scenario is a possible partial description of the world based on what we know and what we are allowed to assume. Consistency means that we do not want to make assumptions that we know are false; this seems like a minimal requirement for rationality.

Definition: If g is a closed formula, an **explanation** of g is a scenario that implies g.

Definition: an **extension** is the set of logical consequences of a maximal (with respect to set inclusion) scenario.

Theorist is an attempt to be a minimalist system. It is an attempt to see how far we can go with a very simple hypothetical reasoning framework. It is also of interest because exactly the same formal definition provides a definition for default reasoning [Poole88a] abductive reasoning (where we want an explanation of an observation in terms of clauses [PGA87, Poole88b]), and also a definition of design (where we want to hypothesize a system which provably fulfils some design requirements [Finger85]).

4 Representational Methodology

4.1 Abduction, Default Reasoning and Design

Different uses of Theorist can be characterised by (a) who chooses the assumptions and (b) whether the goal is known or not. The first considers whether the system is free to choose any hypothesis that it wants or whether "nature" has already chosen the hypothesis and the system has to try to guess the hypothesis chosen. The second is whether the goal is known to be true or whether it is something that has to be determined.

Abduction is where the system knows that the goal (the observation of the world) is true, but it is not free to choose just any hypotheses. Which object is in a scene, or which disease a person has, has already been determined; all we can do is to guess what is in the world based on our observations of the world. We thus consider all explanations of the observations as being possible descriptions of the world. We only need to consider the minimal and least presumptive explanations, as if the set of all explanations is covering, the set of minimal and least presumptive explanations will be as well [Poole89a].

Default reasoning can be seen as where the system does not know whether the goal is true, and is not free to choose any defaults it likes. One appealing framework is to predict something only if is explained even when an adversary chooses the hypotheses [Poole89a], which corresponds to membership in all extensions (which corresponds, propositionally at least, to circumscription [Etherington88]). This is a very sceptical sort of default prediction.

Design can be defined as when the system can choose any hypothesis it wants. For example, a system can choose the components of the design in order to fulfill its design goal, or choose utterances to make in order to achieve a discourse goal. The consistency check is used to rule out impossible designs. All other sets of components that fulfil the goal are possible, and the system can choose the "best design" to suit its goal. Design can be done in an abductive way to try to hypothesise components in order to imply a design goal. Alternatively, design can be done in a default reasoning way to prove a design from goals and any hypotheses we care to choose.

Note that these frameworks are different ways to use the same formal system for different purposes. All ways to use the system may be present in the same system [Poole90].

4.2 Recognition

We have divided the sorts of assumption based reasoning considered here into 3 sorts.[1] To fully define the

[1] Note that learning and scientific theory formation are conspicuous by their absence. This is in order to simplify

theory it remains to specify how such reasoning should be used.

Suppose the problem is a recognition task: given an observation about the world to find out what could be the underlying reality that it corresponds to. This problem can be cast into the hypothetical reasoning framework of Theorist in at least two different ways [Poole88b, Poole89c]:

1. We can treat recognition as an abductive problem, where we find a set of hypotheses that can be added to the knowledge base in order to prove the observations.
2. We can treat this as a prediction problem where the problem is to find what follows from the knowledge base and the observations, perhaps being able to hypothesise defaults.

We can think of recognition as finding the "causes" of the observations. For abduction we have to axiomatise $cause \rightarrow effect$ knowledge. For prediction we have to axiomatise $effect \rightarrow cause$ knowledge. Note that the axiomatization of the knowledge does not depend on the problem domain but rather in the way that the knowledge is to be used.

For the propositional case, suppose $c_1, ..., c_n$ to be the possible causes[2] of symptom s, where $c_1, ..., c_k$ are those causes that always produce s. There are a number of ways this knowledge can be represented:

1. For the abductive systems, we need $c_i \Rightarrow s$ as a fact for $1 \leq i \leq k$ (as we want $\neg s$ to rule out c_i). We need $c_j \Rightarrow s$ to be a possible hypothesis for $k < j \leq n$ (we don't want to rule out c_j by finding out $\neg s$). We also need c_i for $i = 1..n$ to be a possible hypothesis that can be hypothesised if we observe s [Poole90].

   We have to be able to anticipate all possible things that can be observed, as we need to be able to find an explanation of all observations, even observations of normality for which "s is acting normally" may be a reasonable hypothesis.

2. For the prediction representation we have a choice:

(a) We can write the closure of the causes explicitly. Thus we can write the formula

$$s \Rightarrow c_1 \vee ... \vee c_n \qquad (1)$$

so that if s is observed we can conclude that one of the possible causes was responsible. We also need the facts $c_i \Rightarrow s$ for $1 \leq i \leq k$ in order to be able to rule out causes if we know the symptom is not observed. Formula 1 forms an explicit statement of complete knowledge (i.e., that these are all of the possible causes).

(b) We can write down local causal rules

$$s \wedge d_i \Rightarrow c_i$$

for $i = 1..n$ as facts, and have d_i as defaults. Again, we also need the facts $c_i \Rightarrow s$ for $1 \leq i \leq k$. There are a number of problems with such a representation including

- Given s we can conclude the conjunction of the c_i as opposed to the disjunction of the c_i. When we have multiple observations we do not group the causes together in a natural manner (for example, instead of concluding $(cold \wedge exercise) \vee flu$ when we observe $aching\_limbs \wedge sneezing$, we find we can conclude $cold \wedge exercise \wedge flu$. There is, however, no evidence for the conjunction)
- There is a problem pointed out by Pearl [Pearl87a] of cascading causal conclusions with evidential reasoning. For example, from *exercise* concluding *aching_limbs*, and using this as evidence for flu.

There seems to be two solutions to these problems:

(i) To solve the second problem we can add extra preconditions to the rules to ensure the evidential rules is only used if the effect has been concluded by virtue of evidential reasoning and not by causal reasoning [Pearl87a]. It is, however, not so obvious how to use this idea to solve the first problem. There is also a problem that arises when we have both causal and evidential reasons for a particular effect. The preconditions will allow us to hypothesise extra causes for the evidence.

(ii) We can add disabling rules to the knowledge base. In order to get the same answers (in all extensions) as the completion case 2(a), we need to add the rules $i \neq j \wedge d_i \rightarrow \neg d_j$. Then $d_1 \vee ... \vee d_n$ is in all extensions, from which we can derive equation 1. To make sense of these cancellation axioms, the default d_i should be read as "c_i is the primary cause of symptom s" (and we only want one primary cause).

If we are considering what is in any extension, we can still get into the second problem above. From c_1 we can use causal rules to conclude s and then assume d_2 to allow us to conclude c_2. To fix this problem we can add the cancellation axiom $c_i \Rightarrow \neg d_j$ for $i \neq j$. This, however, lets

---

the development of the theory and also is following a common theme in AI that we should try to understand the process of reasoning before we consider the problems of how such reasoning can be learnt.

[2]This notion of causality is very weak, for example one cause may be "it just happened that s", or "the normal state of affairs for s" (in which case we don't really want a deeper explanation of why s occurred). Causality is not to be regarded as anything deep; it is regarded here as a view that is imposed on the world, and is not necessarily intrinsic in the world.

us conclude $\neg c_i$ by assuming d_j. Concluding the negation of causes may not seem like a problem, however it is if some proposition and its negation can be a cause. For example, getting an "A" in a course may be the cause of some actions, and its negation may be the cause of other actions. Peculiar side effects may follow from assuming the negation of *got_A*.

One of the important differences between the first and the second is in the level of detail required of the result. In the abductive case the explanation needs to be at the detail to logically imply the observation. The detail of the explanation is thus determined by the observation. For the second case the level of detail is determined by the knowledge base and not by the observation.

4.3 Hybrid Abduction-Prediction system

One intuitively appealing architecture we have considered [Poole89a, Poole90] is where the different modes of reasoning ar combined, and reasoning proceeds by first abducing causes and then using membership in all extensions to see what is predicted from these causes. The major advantage of this architecture is that we only need $cause \rightarrow effect$ rules, and these same rules can be used for both explanation and prediction.

In order to make this work we need to distinguish two types of possible assumptions; those used for abduction and those used for prediction. In most domains there are possible hypotheses that we want to use for abduction that we do not want to use for prediction.

See [Poole90] for a detailed examination of a representational methodology for this architecture.

4.4 Design-Recognition Duality

Another piece that we have to fit into this jigsaw is our notion of design. I want to argue that there is a duality between the design problem and the recognition problem.

To understand this duality consider a discourse where the speaker is designing her utterances and the hearer is trying to recognise the beliefs or goals underlying the utterance. A useful assumption may be that they share their assumptions.

If the speaker is using assumption-based reasoning to derive an utterance (i.e., he proves an utterance based on assumptions), in order to share the assumptions, the hearer should do abductive reasoning to find what assumptions were needed to imply the utterance.

Suppose, however, the speaker did a form of abductive reasoning where he hypothesised actions (in this case utterances) that would allow him to conclude his goal. In this case to share the forms of assumptions the hearer would need to do prediction. That is, she tries to see what follows from the utterances based on assumptions that the speaker may be using.

If the speaker can choose any assumptions then the best the hearer can do is conclude what is in all extensions. If however the hearer follows some rules to restrict his assumptions, then the hearer can also follow the rules in order to give a more refined sense of prediction (thus effectively broadening the band-width of the communication channel).

What is important about this is that we need both abductive reasoning and prediction style reasoning in each case, but we still have a choice as to which we use for generation and which we use for recognition. What trade-offs are involved is a question we are still investigating [Csinger90].

4.5 Problems with cancellation axioms.

We have considered a number of choices that we can make in our representational methodology, and have discussed some alternatives that do not seem to work for various reasons. When building the representational methodology it is important to keep trying break the system and determining what ideas do not work. Although we have only very limited experience, there are a couple of problems that we have found arise with the naive system presented so far (and many other systems too).

The first has to do with the facilities available to prevent the applicability of defaults. In this discussion we have used a form of cancellation axioms. If we want $c \Rightarrow s$ to be a default that is not applicable under condition e, then we "name" the default [Poole88a], with name d, make d a possible hypothesis, make $d \wedge c \Rightarrow s$ a fact, and make $e \Rightarrow \neg d$ a fact. It is interesting to know that, while this works for simple examples, it runs into problems for larger problems.

As an example, consider representing the defaults "birds fly","emus are birds that don't fly" and "if something looks like an emu it is an emu". Suppose the first default is named $bf(X)$, so we have the axiom $bf(X) \wedge bird(X) \Rightarrow flies(X)$. If we want to conclude that an emu flies, we need to cancel this default for emus and use the axiom $emu(X) \Rightarrow \neg bf(X)$. Using this axiom we can conclude that any individual not known to be an emu is not an emu by assuming bf for that individual. For something that looks like an emu we need to cancel the defaults that lets us conclude the object is not an emu. We thus need the axiom

$$looks\_like\_emu(X) \Rightarrow \neg bf(X)$$

This knowledge base gets the "right" answer for most simple examples, but it doesn't work for the case where some object looks like an emu, is not an emu, but is a bird. In this case we cannot conclude that the object flies. The reason is that we were forced to add the above cancellation axiom which blocks the correct conclusion.

Brewka [Brewka89] gives other arguments based on complexity as to why simple cancellation doesn't

work. He suggested prioritisation of defaults, but unfortunately, no one has suggested a representational methodology to say how to choose the priorities.

Another problem arises when we can derive that there is no individual of a certain sort that is normal in every respect (e.g., when we know that every bird is peculiar in a some way). In this case we get a qualitative version of the lottery paradox. The conjunction of the defaults cannot be used, so membership in all extensions does not let us use any of the defaults [Poole89b].

4.6 Tractability

To test our representational conjecture we don't try to find examples where it works, but rather we try to show that it is false. We could show that it is false by showing that it is incapable of representing the sort of reasoning we need for real problems, by showing that it is not able to be implemented, or by showing that it is inherently inefficient.

Showing Theorist is intractable or undecidable do *not* prove it is useless. These indicate that it is powerful, not that it is inefficient. The property that we would like, is that representing a problem in Theorist does not increase the computational complexity of the problem. We want the ability to solve simple problems simply, while preserving the ability to solve difficult problems.

5 Conclusion

One of the aims of this paper was to convince the reader that Theorist is a natural and commonsense way to reason with defaults. The Theorist research is interesting, I believe for a number of reasons:

It is simple, powerful, and can be motivated in a very natural way.

It can be simply and efficiently implemented[3] [PGA87]. It has been used for many applications.

Exactly the same formal system can be seen as a basis for default reasoning, abductive reasoning and for design. Thus there are independent ways of motivating the same system.

There is a conjecture that the Theorist framework is all that is needed for all forms of reasoning. By showing how this conjecture is false we will have found a principled reason to add more advanced features to our repertoire.

Theorist is probably most important for what it is not. It is not a new logic, it does not need a new semantics, there are no new operators or rules of inference. I have tried to be careful in arguing that we should consider useful ways to use normal logic to build AI programs and applications rather than inventing formalisms that we may not need anyway.

[3] A compiler from Theorist to Prolog is available electronically from the author.

Acknowledgements

This work could not have been done without the ideas, criticism, feedback and support of Randy Goebel, Eric Neufeld, Romas Aleliunas and other members of the (now distributed) Logic Programming and AI group. Thanks to Andrew Csinger for comments on this paper. This research was supported under NSERC grant OPPO044121.

References

[Brewka89] G. Brewka, "Preferred Subtheories: an extended logical framework for default reasoning" *Proc. IJCAI-89*, 1043-1048.

[Csinger90] A. Csinger and D. Poole, "Hypothetical Reasoning and Discourse Structure", forthcoming.

[Etherington88] D. Etherington, *Reasoning with Incomplete Information*, Morgan Kaufmann.

[Finger85] J. Finger and M. Genesereth, "Residue: a deductive approach to design synthesis", Report STAN-CS-85-1035.

[Levesque88] H. Levesque (Ed.), "Taking Issue/Forum: A Critique of Pure Reason", *Computational Intelligence*, 3(3), 149-237.

[Pearl87a] J. Pearl, "Embracing causality in formal reasoning", *Proc. AAAI-87*, 369-373.

[PGA87] D. L. Poole, R. G. Goebel, and R. Aleliunas, "Theorist: a logical reasoning system for defaults and diagnosis", in N. Cercone and G. McCalla (Eds.) *The Knowledge Frontier: Essays in the Representation of Knowledge*, Springer Varlag, 331-352.

[Poole88a] D. Poole, "A Logical Framework for Default Reasoning", *Artificial Intelligence*, 36(1), 27-47.

[Poole88b] D. Poole, "Representing Knowledge for Logic-based diagnosis", *Proc. International Conference on Fifth Generation Computing Systems*, Tokyo, 1282-1289.

[Poole89a] D. Poole, "Explanation and prediction: an architecture for default and abductive reasoning", *Computational Intelligence*, 5(2), 97-110.

[Poole89b] D. Poole, "What the lottery paradox tells us about default reasoning", *Proceedings of the First International Conference on the Principles of Knowledge Representation and Reasoning*, 333-340.

[Poole89c] D. Poole, "Normality and Faults in Logic-based Diagnosis", *Proceedings IJCAI-89*, 1304-1310.

[Poole90] D. Poole, "A Methodology for using a default and abductive reasoning system", to appear *International Journal of Intelligent Systems.*

Incremental, Approximate Planning: Abductive Default Reasoning

Charles Elkan
Department of Computer Science
University of Toronto*

ABSTRACT: This paper presents an abductive strategy for discovering and revising plausible plans. Candidate plans are found quickly by allowing them to depend on unproved assumptions. The formalism used for specifying planning problems makes explicit which antecedents of rules have the status of default conditions, and they are the only ones that may be left unproved, so only plausible plans are produced. Candidate plans are refined incrementally by trying to justify the assumptions on which they depend. The new planning strategy has been implemented, and the first experimental results are encouraging.

1 Introduction

Because of uncertainty and because of the need to respond rapidly to events, the traditional view of planning (deriving from STRIPS [Fikes *et al.*, 1972] and culminating in TWEAK [Chapman, 1987]) must be revised drastically. That much is conventional wisdom nowadays. One point of view is that planning should be replaced by some form of improvisation [Brooks, 1987]. However improvising agents are doomed to choose actions whose optimality is only local. In many domains, goals can only be achieved by forecasting the consequences of actions, and choosing ones whose role in achieving a goal is indirect. Thus traditional planners must be improved, not discarded.

This paper addresses the issue of how to design a planner that is incremental and approximate. An approximate planner is one that can find a plausible candidate plan quickly. An incremental planner is one that can revise its preliminary plan if necessary, when allowed more time.

*For correspondence: Department of Computer Science, University of Toronto, Toronto M5S 1A4, Canada, (416) 978-7797, cpe@ai.toronto.edu.

It is not clear how existing planning strategies can be made approximate and incremental. We therefore first outline a strategy for finding guaranteed plans using a new formalism for describing planning problems, and then show how to extend this guaranteed strategy to make it approximate and incremental.

Our approach draws inspiration from work on abductive reasoning. A plan is an explanation of how a goal is achievable: a sequence of actions along with a proof that the sequence achieves the goal. An explanation is abductive (as opposed to purely deductive) if it depends on assumptions that are not known to be justified. We find approximate plans by allowing their proofs of correctness to depend on unproved assumptions. Our planner is incremental because, given more time, it refines and if necessary changes a candidate plan by trying to justify the assumptions on which the plan depends.

The critical issue in abductive reasoning is to find plausible explanations. Our planning calculus uses a nonmonotonic logic that makes explicit which antecedents of rules have the epistemological status of default conditions. The distinguishing property of a default condition is that it may plausibly be assumed. These antecedents are those that are allowed to be left unjustified in an approximate plan. Concretely, every default condition in the planning calculus expresses either a claim that an achieved property of the world persists in time, or that an unwanted property is not achieved. Thus the approximate planning strategy only proposes reasonable candidate plans.

Sections 2 and 3 below present the formalism for specifying planning problems and the strategy for finding guaranteed plans. In Section 4 the strategy is extended to become approximate and incremental. Section 5 contains experimental results, and finally Section 6 discusses related and future work.

2 The planning formalism

Different formal frameworks for stating planning problems vary widely in the complexity of the problems they can express. Using modal logics or reification, one can reason about multiple agents, about the temporal properties of actions, and about what agents know [Moore, 1985; Konolige, 1986; Cohen and Levesque, to appear in 1990]. On the other hand, the simplest planning problems can be solved by augmented finite state machines [Brooks *et al.*, 1988], whose behaviour can be specified in a propositional logic. The planning problems considered here are intermediate in complexity. They cannot be solved by an agent reacting immediately to its environment, because they require maintaining an internal theory of the world, in order to project the indirect consequences of actions. On the other hand, they involve a single agent, and they do not require reasoning about knowledge or time.

Our nonmonotonic first-order logic for specifying this type of planning problem is called the PERFLOG calculus.[1] The formal aspects of the calculus will be discussed elsewhere; for the purposes of this paper PERFLOG axioms can be understood intuitively as logic program rules, and we shall just use the Yale shooting problem [Hanks and McDermott, 1986] to introduce the calculus by example.

Two "laws of nature" are central. In the following rules, think of S as denoting a state of the world, of A as denoting an action, and of $do(S, A)$ as denoting the state resulting from performing the action A in the initial state S. Finally, think of P as denoting a contingent property that holds in certain states of the world: a fluent.

$$causes(A, S, P) \rightarrow holds(P, do(S, A)) \quad (1)$$

$$holds(P, S) \wedge \neg\, cancels(A, S, P) \rightarrow holds(P, do(S, A)). \quad (2)$$

Rule (1) captures the commonsense notion of causation, and rule (2) expresses the commonsense "law of inertia": a fluent P holds after an action A if it holds before the action, and the action does not cancel the fluent. Note that since in addition to A, one argument of *causes* and of *cancels* is S, the results of an action (that is, the fluents it causes and cancels) may depend on the state in which the action is performed, and not just on which action it is.

[1] PERFLOG is an abbreviation for "performance-oriented perfect model logic": the formal meaning of a set of PERFLOG axioms is its perfect model as defined in [Przymusiński, 1987].

Given rules (1) and (2), a particular planning domain is specified by writing axioms that mention the actions and fluents of the domain, and say which actions cause or cancel which fluents. In the world of the Yale shooting problem, there are three fluents, *loaded*, *alive*, and *dead*, and three actions, *load*, *wait*, and *shoot*. The relationships of these fluents and actions are specified by the following axioms:

$$causes(S, load, loaded) \quad (3)$$

$$holds(loaded, S) \rightarrow causes(shoot, S, dead) \quad (4)$$

$$holds(loaded, S) \rightarrow cancels(shoot, S, alive) \quad (5)$$

$$holds(loaded, S) \rightarrow cancels(shoot, S, loaded). \quad (6)$$

The initial state of the world s_0 is specified by saying which fluents are true in it:

$$holds(alive, s_0). \quad (7)$$

According to the nonmonotonic semantics of PERFLOG collections of rules,

$$holds(dead, do(do(do(s_0, load), wait), shoot))$$

is entailed by rules (1)–(7). The Yale shooting problem is thus solved.

3 Finding guaranteed plans

The previous section showed how to state the relationships between the actions and fluents of a planning domain as a PERFLOG set of axioms. This section describes a strategy for inventing plans using such a set of axioms; the next section extends the strategy to be approximate and incremental.

A PERFLOG set of axioms is a set of general logic program clauses, and the strategy presented here is in fact a general procedure for answering queries against a logic program.

Iterative deepening. The standard PROLOG query-answering strategy is depth-first exploration of the space of potential proofs of the query posed by the user. Depth-first search can be implemented many times more efficiently than other exploration patterns, but it is liable to get lost on infinite paths. Infinite paths can be cut off by imposing a depth bound. The idea of iterative deepening is to repeatedly explore the search space depth first, each time with an increased depth bound [Stickel and Tyson, 1985].

Iterative deepening algorithms differ in how the depth of a node is defined. One depth measure that performs well, called conspiracy depth, is presented in [Elkan,

1989]. Informally, this measure says that a subgoal is unpromising if its truth is only useful in the event that many other subgoals are also true.

Negation-as-failure. Given a negated goal, the negation-as-failure idea is to attempt to prove the un-negated version of the goal. If this attempt succeeds, the negated goal is taken as false; otherwise, the negated goal is taken as true. Negation-as-failure is combined with iterative deepening by limiting the search for a proof of each un-negated notional subgoal. If this search terminates without finding a proof, then the original negated subgoal is taken as true. If a proof of the notional subgoal is found, then the negated subgoal is taken as false. If exploration of the possible proofs of the notional subgoal is cut off by the current depth bound, it remains unknown whether or not the notional subgoal is provable, so for soundness the actual negated subgoal must be taken as false.

Negation-as-failure is only correct on ground negated subgoals, so when a negated subgoal is encountered, it is postponed until finding answers for other subgoals makes it become ground. This process is called freezing [Naish, 1986]. If postponement is not sufficient to ground a negated subgoal, then an auxiliary subgoal is introduced to generate potential answers. This process is called constructive negation [Foo *et al.*, 1988].

The performance of the planning strategy just described could be improved significantly, notably by caching subgoals once they are proved or disproved [Elkan, 1989]. Nevertheless it is already quite usable.

4 Finding plausible plans

This section describes modifications to the strategy of the previous section that make it approximate and incremental. In the same way that the guaranteed planning strategy is in fact a general query-answering procedure, the incremental planning strategy is really a general procedure for forming and revising plausible explanations using a default theory.

Any planning strategy that produces plans relying on unproved assumptions is *ipso facto* unsound, but by its incremental nature the strategy below tends to soundness: with more time, candidate plans are either proved to be valid, or changed.

Approximation. The idea behind finding approximate plans is simple: an explanation is approximate if it depends on unproved assumptions. Strategies for forming approximate explanations can be distinguished according to the class of approximate explanations that each may generate. One way to define a class of approximate explanations is to fix a certain class of subgoals as the only ones that may be taken as assumptions. Looking at the PERFLOG formalism, there is an obvious choice of what subgoals to allow to be assumptions. Negated subgoals have the epistemological status of default conditions: the nonmonotonic semantics makes them true unless they are forced to be false. It is reasonable to assume that a default condition is true unless it it is provably false.

There is a second, procedural, reason to allow negated subgoals to be assumed, but not positive subgoals. Without constructive negation, negated subgoals can only be answered true or false. Negation-as-failure never provides an answer substitution for a negated subgoal. Therefore unproved negated subgoals in an explanation never leave "holes" in the answer substitution induced by the explanation. Concretely, a plan whose correctness proof depends on unproved default conditions will never change because those defaults are proved to hold.

Incrementality. An approximate explanation can be refined by trying to prove the assumptions it depends on. If an assumption is proved, the explanation thereby becomes "less approximate". As just mentioned, proving an assumption never causes a plan to change. On the other hand, if an assumption is disproved, the approximate plan is thereby revealed to be invalid, and it is necessary to search for a different plan.

Here are the details of the modifications made to the planning strategy of the previous section. When a negated subgoal becomes ground, the proof of its notional positive counterpart is attempted. If this attempt succeeds or fails within the current depth bound, the negated subgoal is taken as false or true, respectively, as before. However, if the depth bound is reached during the attempted proof, then the negated subgoal is given the status of an assumption.

Initially any negated subgoal is allowed to be assumed. Each iteration of iterative deepening takes place with an increased depth bound. For each particular (solvable) planning problem, there is a certain minimum depth bound at which one or more approximate plans can first be found. Each of these first approximate plans depends on a certain set of assumptions. In later iterations, only subsets of these sets are allowed to be assumed. This restriction has the effect of concentrating attention on either refining the already discovered approximate plans, or finding new approximate plans that depend on fewer assumptions.

5 Experimental results

Implementing the planning strategies described above is straightforward, because the PERFLOG calculus is based on definite clauses. In general, it is insufficiently realized how efficiently logics based on definite clauses, both monotonic and nonmonotonic, can be implemented. The state of the art in PROLOG implementation is about nine RISC cycles per logical inference [Mills, 1989]. Any PERFLOG theory could be compiled into a specialized incremental planner running at a comparable speed.

The experiment reported here uses a classical planning domain: a lion and a Christian in a stadium. The goal is for the lion to eat the Christian. Initially the lion is in its cage with its trainer, and the Christian is in the arena. The lion can jump from the cage into the arena only if it has eaten the trainer. The lion eats a person by pouncing, but it cannot pounce while it is already eating. The following PERFLOG theory describes this domain formally.

```
%
% rules for how the world evolves

holds(P,do(S,A)) :-
        causes(A,S,P).
holds(P,do(S,A)) :-
        holds(P,S), not(cancels(A,S,P)).

%
% the effects of actions

causes(pounce(lion,X),S,eats(lion,X)) :-
        can(S,pounce(lion,X)).
can(pounce(X,Y),S) :-
        holds(in(X,L),S), holds(in(Y,L),S),
        not(call(X = Y)),
        not(Z,holds(eats(X,Z),S)).
causes(jump(X),S,in(X,arena)) :-
        can(jump(X),S), holds(in(X,cage),S).
can(jump(lion),S) :-
        holds(eats(lion,trainer),S).
cancels(drop(X,Y),S,eats(X,Y)) :-
        can(drop(X,Y),S).
can(drop(X,Y),S) :-
        holds(eats(X,Y),S).
holds(in(X,H),S) :-
        holds(eats(lion,X),S), holds(in(lion,H),S).

%
% the initial state of the world

holds(in(christian,arena),s0).
holds(in(lion,cage),s0).
holds(in(trainer,cage),s0).
```

Using the guaranteed planning strategy of Section 3, the query `holds(eats(lion,christian),P)?` is first solved with conspiracy depth bound 19, in 4.75 seconds.[2] The plan found is

```
P = do(do(do(do(s0,pounce(lion,trainer)),
                   jump(lion)),
                   drop(lion,trainer)),
                   pounce(lion,christian)).
```

Using the approximate planning strategy of Section 4, the same query is solvable in 0.17 seconds, with conspiracy depth bound 17. The candidate plan found is

```
P = do(do(do(s0,pounce(lion,trainer)),
                jump(lion)),
                pounce(lion,christian)).
```

This plan depends on the assumption that no `Z` exists such that

```
holds(eats(lion,Z),do(do(s0,pounce(lion,trainer)),
                                jump(lion))).
```

Although the assumption is false and the plan is not correct, it is plausible. Note also that the first two actions it prescribes are the same as those of the correct plan: the approximate plan is an excellent guide to immediate action.

6 Discussion

The work reported here ties together ideas from a number of different research areas.

Approximate planning. From a knowledge-level point of view, the strategy for finding plausible plans is searching in an abstraction space where the available actions are the same as in the base space, but they are stripped of their difficult-to-check preconditions. Compared to other abstraction spaces [Knoblock, 1989], this space has the advantage that the execution of a plan invented using it can be initiated without further elaboration, if immediate action is necessary.

Incremental planning. An incremental approximate planner is an "anytime algorithm" for planning in the sense of [Dean and Boddy, 1988]. Anytime planning algorithms have been proposed before, but not for problems of the traditional type treated in this paper. For example, the real-time route planner of [Korf, 1987] is a heuristic graph search algorithm, and the route improvement algorithm of [Boddy and Dean, 1989] relies on an initial plan that is guaranteed to be correct.

[2] All times are for an implementation in CProlog, running on a Silicon Graphics machine rated at 20 MIPS.

Abductive reasoning. Abduction mechanisms have been investigated a great deal for the task of plan recognition, not so much for the task of inventing plans, and not at all for the task of inventing plausible plans. These three different tasks lead to different choices of what facts may be assumed. In the work of [Shanahan, 1989] for example, properties of the initial state of the world may be assumed. In our work, the facts that may be assumed say either that an established property of the world persists, or that an unestablished property does not hold.

Directions for future work. One important problem is to quantify how an approximate plan is improved by allowing more time for its refinement. Another problem is to find a planning strategy that is focused as well as approximate and incremental. A focused strategy would be one that concentrated preferentially on finding the first step in a plan–what to do *next*.

References

[Boddy and Dean, 1989] Mark Boddy and Thomas Dean. Solving time-dependent planning problems. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 979–984, August 1989.

[Brooks *et al.*, 1988] Rodney A. Brooks, Jonathan H. Connell, and Peter Ning. Herbert: A second generation mobile robot. MIT AI Memo 1016, January 1988.

[Brooks, 1987] Rodney A. Brooks. Planning is just a way of avoiding figuring out what to do next. Technical Report 303, Artificial Intelligence Laboratory, MIT, September 1987.

[Chapman, 1987] David Chapman. Planning for conjunctive goals. *Artificial Intelligence*, 32:333–377, 1987.

[Cohen and Levesque, to appear in 1990] Philip R. Cohen and Hector J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, to appear in 1990.

[Dean and Boddy, 1988] Thomas Dean and Mark Boddy. An analysis of time-dependent planning. In *Proceedings of the National Conference on Artificial Intelligence*, pages 49–54, August 1988.

[Elkan, 1989] Charles Elkan. Conspiracy numbers and caching for searching and/or trees and theorem-proving. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 341–346, August 1989.

[Fikes *et al.*, 1972] Richard E. Fikes, Peter E. Hart, and Nils J. Nilsson. Learning and executing generalized robot plans. *Artificial Intelligence*, 3:251–288, 1972.

[Foo *et al.*, 1988] Norman Y. Foo, Anand S. Rao, Andrew Taylor, and Adrian Walker. Deduced relevant types and constructive negation. In Kenneth Bowen and Robert Kowalski, editors, *Fifth International Conference and Symposium on Logic Programming*, volume 1, pages 126–139, Seattle, Washington, August 1988. MIT Press.

[Hanks and McDermott, 1986] Steve Hanks and Drew McDermott. Default reasoning, nonmonotonic logics, and the frame problem. In *Proceedings of the National Conference on Artificial Intelligence*, pages 328–333, August 1986.

[Knoblock, 1989] Craig A. Knoblock. Learning hierarchies of abstraction spaces. In *Proceedings of the Sixth International Workshop on Machine Learning*, pages 241–245. Morgan Kaufmann Publishers, Inc., 1989.

[Konolige, 1986] Kurt Konolige. *A Deduction Model of Belief.* Pitman, 1986.

[Korf, 1987] Richard E. Korf. Real-time path planning. In *Proceedings of the DARPA Knowledge-Based Planning Workshop*, 1987.

[Mills, 1989] Jonathan W. Mills. A pipelined architecture for logic programming with a complex but single-cycle instruction set. In *Proceedings of the IEEE First International Tools for AI Workshop*, September 1989.

[Moore, 1985] Robert C. Moore. *A Formal Theory of Knowledge and Action.* Ablex, 1985.

[Naish, 1986] Lee Naish. *Negation and Control in* PROLOG. Number 238 in Lecture Notes in Computer Science. Springer Verlag, 1986.

[Przymusiński, 1987] Teodor C. Przymusiński. On the declarative semantics of stratified deductive databases and logic programs. In Jack Minker, editor, *Foundations of Deductive Databases and Logic Programming*, pages 193–216, Los Altos, California, 1987. Morgan Kaufmann Publishers, Inc.

[Shanahan, 1989] Murray Shanahan. Prediction is deduction but explanation is abduction. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 1055–1060, 1989.

[Stickel and Tyson, 1985] Mark E. Stickel and W. M. Tyson. An analysis of consecutively bounded depth-first search with applications in automated deduction. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pages 1073–1075, August 1985.

Goal-Directed Diagnosis of Expectation Failures

Bruce Krulwich Lawrence Birnbaum Gregg Collins

Northwestern University
Institute for the Learning Sciences and
Department of Electrical Engineering and Computer Science
Evanston, Illinois

ABSTRACT

Expectation failure diagnosis involves explaining why a faulty belief was inferred. Typical approaches to failure diagnosis have taken this problem to be an independent task, ignoring the goals of the system desiring the diagnosis. This paper discusses the effect that such higher-level goals can have on the process of failure diagnosis, and suggests that failure-driven learning should be viewed explicitly as a goal-directed planning task.

Expectation failure diagnosis

Expectation failure diagnosis is the problem of determining why a problem-solving system has inferred a faulty belief, usually in the service of fixing the system so that the same mistakes won't be repeated in the future (see, e.g., [Sussman, 1975; Schank, 1982]). The most common approach to this problem is to start from a description of the failure and perform a more or less undirected search for a causal chain showing why it occurred. In addition a number of diagnosis systems have been proposed that utilize knowledge of the assumptions that the system used in initially inferring the belief [deKleer, 1987; Simmons, 1988; Chien, 1989; Collins, Birnbaum, and Krulwich, 1989]. Such knowledge can constrain the search for a failure explanation to beliefs and inferences that were relevant to the original faulty belief.

Very few systems, however, take the higher-level goals of the system into account in performing explanation, and those that do (e.g., [Kedar-Cabelli, 1987; Leake, 1988]) treat this as a seperate step that occurs before or after the explanation component is invoked. We will see that the actual *process* of constructing a functionally useful explanation will be affected by these goals. Further, we will propose that the process of diagnosing and correcting failures can best be viewed as a *planning* task, since a consideration of the impact of goals on the explanation process leads to a concern with standard planning problems such as goal interaction.

To understand the effect of higher-level goals on the diagnosis process, consider the situation of a person who decided to leave his car outside on a snowy night when the temperature went below zero. Upon leaving his house in the morning, he finds that the car won't start. A skeletal explanation for this problem is that the car isn't starting because it was left outside overnight on a cold night, and cold weather is bad for cars. Without elaborating the explanation further, it can be seen that, if the explanation is correct, the problem can be prevented in the future by not leaving the car out on cold nights, e.g., by garaging it. If the person is concerned only with making sure that the car will work in the future, this plan will suffice to meet his needs, and no further explanation is necessary.

If, on the other hand, it is crucial that the car be restored to working condition immediately, more explanatory effort will be required. In particular, the explanation will have to be elaborated far enough to pinpoint a problem that can be quickly corrected, such as a run-down battery or a frozen gas line. To generate such an explanation, the person will have to search through a large space of causal knowledge about cars, involving qualitative reasoning about the car's engine, quantitative knowledge of fluid freezing points, and electrical potentials, and knowledge of the precise weather conditions This search is potentially very expensive.

Most aspects of a car are not, however, accessible to our person, assuming that he is not an auto mechanic. Thus, his aim in elaborating the explanation should not be to search the entire space, but rather to determine as quickly as possible whether what is wrong is something that he is able to fix on his own. As soon as it becomes clear that the problem is not going to be fixable, further effort expended on explanation is useless as far as the goal of getting the car going is concerned. A good strategy might, therefore, be to consider whether the problem can be fixed by adding a fluid, recharging the battery, or getting the car rolling, and, if not, suspending further effort and calling a mechanic.

There are several points made by this example. First, the extent to which an explanation must be elaborated will depend upon whether it is currently sufficient to fulfill the planner's overall goal in carrying out the explanation process. Explantions aimed at suggesting plans of action, for example, can be halted as soon as a workable plan is found. Second, the order in which the explainer

searches the space of possible explanations will critically depend upon the goals being addressed. In particular, the planner should focus effort on determining quickly whether the result of the explanation is likely to be useful for the explainer's purposes, or, as in the case of a mechanical explanation of a car problem that does not suggest a fix, will represent a wasted effort. Third, decisions regarding when the explainer should stop explaining and what parts of the explanation it should elaborate first cannot be made *a priori*, but must be made in the course of explaining the failure, since the decision depends upon the precise nature of the explanation suggested. In short, this example demonstrates the need for explicit consideration of higher-level goals in the process of diagnosis.

A fourth point, which is not our main thrust in this paper, but which arises from our analysis, is that an explanation process can take full advantage of the constraints offered by the consideration of the explainer's goals only if the explanation is generated hierarchically. In our example, for instance, it is critical that the explainer generates the explanation that the car is the victim of cold weather before it attempts to elaborate the precise causal mechanism through which the weather affected the car. An explanation process that tried to proceed at a single, predefined level of granularity would risk missing the fact that, for some purposes, the explanation blaming cold weather is good enough.

Let's look now at an example within the domain of our system, which learns strategic concepts from expectation failures that arise in plan execution, in the domain of chess [Krulwich, Collins, and Birnbaum, 1989]. Our system constructs its explanations by searching through explicit justification structures that are maintained for its beliefs by examining these justifications in light of a description of the failure that occurred. It is designed to start out playing chess at a novice level and improve by learning strategic concepts. Suppose that our system has a set of threat detectors that cannot detect *en passant* pawn captures [Birnbaum, Collins, and Krulwich, 1989]. These threats, which are unknown to many novice chess players, involve a pawn that has just made a two-square forward move being captured by another pawn moving (diagonally) into the square that the first pawn skipped over. Such a system may have a pawn which it expects to be safe from attack even though it is in fact susceptible to an *en passant* capture. The justification for the expectation that the piece is safe from attack would be that there is believed to be no threat against the pawn, which would in turn be justified by the fact that none of the system's threat detectors signal a threat against it. The expectation that the piece is safe from attack will fail if the opponent takes the piece using an *en passant* capture, and the system will need to explain why this expectation failed. A skeletal explanation of the failure is simply that the computer advanced the pawn two squares, thinking that there were no threats against it, while the opponent was able to make a move with his pawn that captured it. The computer could decide that it's sufficient to merely satisfy the higher-level goal of ensuring that pawns will not be taken in this manner in the future. For this goal, the skeletal explanation will suffice as it stands, because the fact that the computer's advancing its pawn two squares enabled the computer's move can lead the computer to decide never to advance its pawn two squares. On the other hand, the computer could decide to elaborate the explanation further. To do this completely involves explaining why the computer advanced the pawn, why the threat detection mechanism failed to signal a threat, and how the opponent was able to execute the undetected threat. If the computer's goal is specifically to prevent the incorrect expectation from being made in the future, the system should focus on exactly why its threat detection mechanism didn't signal a threat. This will lead to the explanation that there did not exist a threat detector that could detect the threat that the opponent made.

Each of these two explanations will be used to learn from the failure, and the way in which the system will avoid making the mistake after learning will depend on the goal that was assumed by the diagnosis mechanism. If the system sufficed to ensure in the simplest way that pawns are not taken in this manner in the future, the explanation will be that the computer had advanced its pawn and the opponent had captured it. This explanation could lead to the computer's decision never to advance its pawns two squares at a time. Alternatively, this explanation could lead the computer to modify the plan it was in the process of executing to include a counterplan that eleminates the threat. This would only handle *en passant* threats when the computer is executing the same plan, but this is the cost of not performing the detailed diagnosis of the system's detection mechanism. If, however, the system decided to pursue the general case of not expecting the pawn to be safe in similar situations the explanation will be that there did not exist a threat detector that detected the opponent's move. This explanation will lead the system to add a new threat detector for *en passant* captures.

Several of the points made about the car example should be clear from this example as well. First, a hierarchical diagnosis mechanism is needed so that the system can focus on aspects of the explanation that are relevant to the active higher-level goal. Second, the decisions regarding which aspects of the explanation should be elaborated must be made with respect to the higher-level goals of the system. If the goal is not to expect the pawn to be safe, it's irrelevant why the computer advanced the pawn in the first place, while if the goal is to ensure the pawn's safety in the future it may be irrelevant why the computer was unable to detect the threat. Third, the constraints of the goals on the explanation can only be determined in the process of diagnosis, because they depend on the usefulness of the specific components of the

skeletal explanation.

Goal-directed diagnosis

Let's now consider a diagnosis mechanism that is capable of performing these diagnoses. This mechanism will explain failures in two steps. The first step is to develop a *skeletal explanation* of the failure which will be a deductively correct explanation but will lack explanations for any of the details. In the car example above this skeletal explanation was *the car isn't starting because it was left outside overnight on a cold night and cold weather is bad for cars.* In the example of *en passant* captures, the skeletal explanation was *the computer advanced its pawn two squares, having not detected any threats against it, and the opponent was able to make a move with his pawn that captured it.* These skeletal explanations do in fact explain their respective failures, but they are at too high a level of granularity to be useful in most situations. The second step in the diagnosis, therefore, is to elaborate the aspects of the skeletal explanations that are considered important for the higher-level goals of the system. This is achieved using *diagnosis methods*, which give a method of furthering the explanation for a given aspect of an explanation and a higher-level goal. This mechanism makes the simplifying assumption that the higher-level goals are available *a priori* and can be used to direct the search for an explanation. Components of a skeletal explanation are maintained along with their own subgoals, which reflect how they relate to the failure being explained. Each component of the top-level skeletal explanation is tagged with the same goal as the higher-level goal given to the diagnosis system. As the explanation is elaborated the lower level aspects of the explanation may have different goals, reflecting how they relate to the failure.

To demonstrate the workings of such a diagnosis system, consider the first of the possible higher-level goals in the *en passant* example:

- Don't expect the pawn to be safe in future similar situations

The system should realize that the most important aspect of the skeletal explanation to elaborate is that the threat detection mechanism didn't detect any threats against the pawn. Its goal in explaining why this was the case is to make it untrue in the future, because this will lead to the higher-level goal's being achieved. This subgoal is equivalent to the subgoal to have it be true in similar future situations that the threat detection mechanism detects a threat. The diagnosis mechanism should then use the explicit justification structures to elaborate the statement *the detection mechanism didn't detect a threat* into *no threat detectors detected a threat.* The goal in explaining this is also elaborated, becoming the goal to have a threat detector detect a threat in similar future situations. This aspect of the skeletal explanation does not have to be elaborated any further, because it can be directly used to achieve the goal. The learning component of the system will handle this by taking the statement "there does not exist a threat detector that detects a threat in this situation" as a command to add a threat detector.

- **Explanation aspect:** Negation referring to the system's decision-making
  Goal: Make the implied belief not be inferred in the future
  Method: Explain why the aspect was true, with the goal of having it not be true in the future
- **Explanation aspect:** Negated bounded existentially quantified expression
  Goal: Make the bounded existentially quantified expression true
  Method: Done, with aspect as instruction to add to quantified-over set

Figure 1: Explanation methods for higher-level goal *prevent inference*

To complete this failure diagnosis, the system still has to explain exactly how the opponent moved his pawn to capture the computer's. The diagnosis system should attempt to elaborate on this aspect of the explanation, but it will be unable to do so because the system doesn't know about the *en passant* move used by the opponent. At this point it can try to use general notions of move enablement along with standard explanation-based generalization techniques [DeJong and Mooney, 1986; Mitchell, Keller, and Kedar-Cabelli, 1986], but it may have to ask the user for help in correctly characterizing the move.

In diagnosing the failure with respect to this higher-level goal, the system used the diagnosis methods shown in figure 1. Now consider the second of the possible higher-level goals in the example:

- Ensure that the pawn will be safe in future similar situations

The system should realize here that the aspect of the explanation that says *the computer advanced its pawn two squares* is under the control of the computer, and that if it is avoided in the future the opponent will not be able to make the move that it did. The learning system should realize from this that if it avoids this being the case in the future, that is, if it never advances a pawn two squares, the opponent will not be able to capture pieces in this way. This diagnosis method is shown in figure 2.

We saw in the examples above that particular aspects of a skeletal explanation are likely to have several methods to elaborate on them, each of which is applicable in a different situation. The potential complexity of choosing among and combining different diagnosis methods

- **Explanation aspect:** Proposition referring to the computer's actions
  Goal: Make the implied belief be true in the future
  Method: Done, with the aspect as something to be avoided in the future

Figure 2: Explanation methods for higher-level goal *make inference true*

has led us to propose viewing goal-directed diagnosis as a *planning* process, where the diagnosis module will be given an expectation failure (along with its associated justification) and a higher-level goal, and will decompose this goal to get a diagnosis process that will construct an explanation useful for achieving the given goal[1]. This view of learning as planning is predicated on the fact that typical planning issues, such as interactions between subgoals and choice of alternative goal-satisfaction methods, will arise when an explanation system is given a diagnosis goal and a set of possible methods for explaining the failure given the goal. Viewing diagnosis in this way necessitates explicit representation of diagnosis goals and a theory about the explanations that will be functionally useful in achieving each goal. Within a system of goal-directed failure diagnosis, learning from a failure involves the following steps: (1) Detecting the failure and determining the goal of learning; (2) Explaining the failure in light of the goal; (3) Using the explanation to derive a modification of the system; and (4) Instantiating the modification to learn from the failure.

In the *en passant* example given above, this involves first determining that the expectation that the piece was safe failed, and that the system has a goal of not expecting it to be safe in the future. Second, the system should spawn a goal to explain the pawn's not being safe with the goal of not making the inference in the future. This explanation should be that there did not exist a threat detection rule that detected the threat of the opponent's pawn again the computer's pawn. Third, this explanation gives a fix of adding a threat detector to detect the threat, and instantiating this fix adds the detector.

Conclusion

We have illustrated the effect that goals have on the process of diagnosing expectation failures, and the consequent need to explicitly represent and reason about these goals in a complete model of explanation. We have proposed viewing failure diagnosis and repair as a planning task. Future work will determine the effect this will have on other areas of explanation and learning.

[1]This is different from Hunter's *knowledge acquisition planning* [Hunter, 1989], which deals with a system's planning globally the types of things that it needs to learn.

Acknowledgments: .. would like to thank Michael Freed, Eric Jones, Alex Kass, Smadar Kedar, and Louise Pryor for discussions on the ideas presented here and comments on this paper. This work was funded in part by the Office of Naval Research under contract N00014-89-J-3217. The Institute for the Learning Sciences was established in 1989 with the support of The Arthur Andersen Worldwide Organization.

References

Birnbaum, L., Collins, G., and Krulwich, B. 1989. Issues in the justification-based diagnosis of planning failures. *Proceedings of the Sixth International Workshop on Machine Learning*, Ithaca, NY, pages 194-196.

Chien, S. 1989. Using and refining simplifications. Explanation-based learning of plans in intractable domains. *Proceedings of the Eleventh IJCAI*, Detroit, MI, pages 590-595.

Collins, G., Birnbaum, L., and Krulwich, B. 1989. An adaptive model of decision-making in planning. *Proceedings of the Eleventh IJCAI*, Detroit, MI, pages 511-516.

DeJong, G., and Mooney, R. 1986. Explanation-based learning: An alternative view. *Machine Learning*, vol. 1, pages 145-176

deKleer, and Willams, B.C. 1987. Diagnosing multiple faults. *Artificial Intelligence* **32**, pages 97-130.

Hunter, L. 1989. Knowledge acquisition planning: Results and prospects. *Proceedings of the Sixth International Workshop on Machine Learning*, Ithaca, NY, pages 61-65.

Kedar-Cabelli, S. 1987. Formulating concepts according to purpose. *Proceedings of the 1987 AAAI Conference*, Seattle, WA, pages 477-481.

Krulwich, B., Collins, G., and Birnbaum, L. 1989. Improving decision-making on the basis of experience. *Proceedings of the Sixth International Workshop on Machine Learning*, Ithaca, NY, pages 55-57.

Leake, D. 1988. Evaluating explanations. *Proceedings of the 1988 AAAI Conference*, St. Paul, MN, pages 251-255.

Mitchell, T., Keller, R., and Kedar-Cabelli, S. 1986. Explanation-based generalization: A unifying view. *Machine Learning*, vol. 1, pages 47-80

Schank, R. 1982. *Dynamic Memory: A Theory of Reminding and Learning in Computers and People*. Cambridge University Press, Cambridge, England.

Simmons, R. 1988. A theory of debugging plans and interpretations. *Proceedings of the 1988 AAAI Conference*, St. Paul, MN, pages 94-99.

Sussman, G. 1975. *A Computer Model of Skill Acquisition*. American Elsevier, New York

Parsimonious Covering Theory and Non-Diagnostic Tasks

James A. Reggia
Yun Peng
Department of Computer Science,
University of Maryland
College Park, MD 20742

Parsimonious covering theory (PCT) is a mathematical model of abductive reasoning for diagnostic problem-solving [Reggia et al, 1983, 1985; Peng & Reggia, 1990]. It provides a formal, application independent theory of the underlying causal knowledge and the reasoning processes involved in diagnostic inference, as well as criteria for plausibility (coherence, acceptability) of explanatory hypotheses. This paper begins to examine the extent to which the principles and results of PCT, originally formulated for diagnostic reasoning, can be applied to non-diagnostic tasks. A brief introduction and summary of PCT is given first. Then, PCT is compared to a theory of explanatory coherence in abduction and related to various aspects of natural language processing.

Parsimonious Covering Theory (PCT)

In the simplest form of PCT there is a set of disorders D and a set of manifestations ("symptoms") M. For each disorder d_i, there is a connection (association) between d_i and each manifestation m_j that can be caused by d_i. A subset of M, denoted M^+, represents the set of all manifestations known to be present. A set of disorders D_I is called a *cover* of the given M^+ when the disorders in D_I can cause all of the manifestations in M^+. A set of disorders D_I is an *explanatory hypothesis* if 1) D_I is a cover of M^+, and 2) D_I is parsimonious. A difficult issue in diagnostic reasoning theories in general, including PCT, has been precisely defining what is meant by the "best", "most plausible", "simplest" or "most parsimonious" explanation for a given set of facts [deKleer and Williams, 1986; Josephson et al, 1987; Pople, 1973; Peng and Reggia, 1987; Reggia et al, 1983, 1985; Reiter, 1987]. Previous notions of plausibility have largely been based on subjective criteria; we consider two of these here.

An early criterion of plausibility used in PCT and by others is called *minimal cardinality*: explanatory hypotheses with the fewest number of hypothesized components are preferable. In applying PCT to specific diagnostic problems, it quickly becomes evident that minimal cardinality is an inadequate measure of plausibility. For example, in medical diagnosis two common diseases may together be more plausible than a single rare disease in explaining a given set of symptoms [Reggia et al, 1985], and in electronic diagnosis analogous examples exist [Reiter, 1987]. For this reason, PCT as well as other models of diagnostic inference have adopted a more relaxed criterion of plausibility which we call *irredundancy*: a set of disorders D_I that covers (causes all of) the manifestations in M^+ is irredundant if it has no proper subsets which also cover M^+. Although this criteria does not directly favor the smallest set of propositions, irredundancy is a preferable criterion because it handles cases like the medical and electronics examples referenced above while still constraining the number of disorders in an hypothesis. However, irredundancy has the problem that in applications it may identify many implausible hypotheses as well as the plausible ones, and as indicated below, may in some cases still fail to identify the most reasonable hypothesis.

These criteria, used in most theories of explanatory plausibility, including PCT, are *subjective*. An important issue is whether one might devise *objective* measures of plausibility and then ask under what conditions various subjective criteria would work or fail according to the objective

criteria. Recently, we generalized Bayes' Theorem to apply to diagnostic problems formulated in PCT [Peng and Reggia, 1987]. Each disorder d_i is associated with its prior probability p_i. Each causal link is associated with a number c_{ij}, the causal strength from d_i to m_j representing how frequently d_i causes m_j. Under assumptions less restrictive than those traditionally made with Bayesian classification, the relative likelihood $L(D_l,M^+)$ of any potential explanatory hypothesis D_l given the presence of M^+ can be calculated using relevant p_i and c_{ij} values. $L(D_l, M^+)$ can be proven to differ from the posterior probability $P(D_l|M^+)$ by only a constant. Using the objective, albeit limited, measure $L(D_l,M^+)$, one can ask under what conditions various plausibility criteria such as minimal cardinality, irredundancy, and others would be guaranteed to identify the most probable hypothesis.

Analytical treatment of this question leads to a number of interesting results [Peng and Reggia, 1987]. For example, minimal cardinality is only an appropriate criterion when, for all disorders d_i, the prior probabilities are very small and about equal, and the c_{ij} are fairly large in general. Otherwise, it may be that the most probable explanation does not have minimal cardinality, supporting the conclusion above that counting is not sufficient. In fact, there are situations where the most probable explanation does not even satisfy irredundancy.

With this brief background, we now turn to the issue of whether the principles and results in PCT can be adapted for use in non-diagnostic problems. The general area of theory formation is considered first, followed by some aspects of natural language processing.

Theory of Explanatory Coherence (TEC)

The Theory of Explanatory Coherence (TEC) is a general framework for considering the plausibility of explanatory hypotheses [Thagard, 1989]. TEC is intended to apply to both scientific reasoning and "reasoning in everyday life", which certainly includes diagnostic inferences. Its foundations are a set of seven heuristic principles that describe the "coherence" and "acceptability" of explanatory hypotheses. TEC differs from PCT in its informal (as opposed to mathematical) formulation and TEC's boader orientation towards general abductive reasoning rather than diagnosis. Roughly speaking, asserting the presence/absence of manifestation m_j or disorder d_i in PCT corresponds to a proposition in TEC, and a parsimonious cover represents specification of the function which defines system coherence (TEC Principle 7). Because of its restricted applicability to diagnostic problem-solving, PCT does not address some issues of TEC (e.g., analogy). However, like TEC, PCT precisely defines the notion of explanatory hypotheses and what makes them plausible, has been applied to specific applications, and has been formulated as a connectionist model [Peng & Reggia, 1989; Wald et al, 1989].

We now focus on comparing some of the fundamental principles or assumptions underlying TEC and PCT. The first observation is that, to the extent they can be compared, these two independently-developed theories are in broad and general agreement. Starting from elementary hypothesis elements, both theories are concerned with the construction of composite hypotheses that can account for observed data. Both theories give priority to observational data, and both adopt some notion of parsimony in judging the plausibility of competing hypotheses. This broad, top-level correspondence between the basic principles in TEC and PCT allows one to conclude that, whatever the impreciseness in our current definition of abduction, there is at least the beginnings of a concensus on some of the fundamental properties of an abductive inference system.

However, the differences between TEC and PCT are more interesting. Two of them are briefly considered here. Consider first one of the central principles of TEC:

> TEC Principle 2c:
> "If $P_1, \ldots, P_m$ explain Q, then . . .
> the degree of coherence is inversely
> proportional to the number of
> propositions $P_1, \ldots, P_m$."

TEC's measure "degree of coherence" is related to the notions of plausibility and probability of explanatory hypotheses in PCT. As noted above, our experience with PCT and diagnosis suggests that counting propositions (TEC Principle 2c) can be an inadequate measure of "coherence" or plausibility.

It has been pointed out [Thagard, 1989] correctly that in some nondiagnostic domains the probabilities do not exist. They do not really exist

in most diagnostic applications either. However, since TEC and PCT are intended to be theories that encompass diagnostic reasoning, they cannot ignore measures of likelihood that go beyond counting, be they numeric probabilities *or other nonnumeric, subjective measures*. Some measure of "prior plausibility" or "intrinsic merit" and "conditional plausibility" of causation is essential in diagnosis, and seems to be just as important in scientific and legal reasoning as well. Basing coherence on counting propositions as in TEC Principle 2c would therefore need revision.

Another TEC principle states that if many relevant observations are unexplained, then the coherence of a hypothesis component is reduced. Specifically,

> TEC Principle 6b:
> "If many results of relevant experimental observations are unexplained, then the acceptability of a proposition P that explains only a few of them is reduced."

This seems to imply that the plausibility of a hypothesis element increases when new evidence supporting it is given. However, a consequence of the Bayesian analysis of PCT [Peng & Reggia, 1987] is the conclusion that this apparently reasonable heuristic is not always correct. A new observation may sometimes cause a decrease in the likelihood of a hypothesis component that can cause/explain that observation if that observation supports a rival hypothesis element more strongly at the same time.

It can be concluded that the general heuristic principles of TEC and the diagnosis-specific framework of PCT are in agreement about central issues. However, the experiences with PCT suggests that some of the details of TEC will need to evolve further if it is to serve as a general theory of abduction.

Natural Language Processing

Recently, a growing number of AI researchers have been working with the assumption that abductive inference underlies natural language processing. For example, natural language processing involves context-sensitive disambiguation of word senses and inferences about plausible explanations at a low level (e.g., garden path sentences, ellipses, and anaphora resolution). Similar examples exist for high level natural language processing such as inferring the plans of the participants in a dialog.

As an example, an analogy between diagnostic problem-solving as formulated in PCT and word sense disambiguation in natural language processing can be identified:

PCT	Word-Sense Disambiguation
manifestation	word
disorder	sense
causal relation	word-sense assoc.

In terms of the knowledge used, both tasks involve the use of associative knowledge. Like disorders and manifestations associated by causal relations in diagnostic problems, natural language processing involves associations between linguistic entities (e.g., word senses) and their manifestations (e.g., words). Such associative knowledge in both tasks is ambiguous. Similar to a manifestation having multiple possible causative disorders, a word (e.g., *fly*) may also have multiple possible senses or meanings (e.g., small insect, baseball hit high in the air, perform a task rapidly, etc.). The similarity also exists in problem-solving. Like an explanation for a set M^+ of present manifestations, the "meaning" of a sequence W^+ of words as a multiple-component hypothesis must be *constructed* from individual elementary semantic concepts such as word senses. Disambiguation of a word is context-sensitive, and thus a parsimony principle may play a role in this disambiguation. Probabilistic knowledge about the uncertainty of associations and about the average frequency of occurrence of entities may also play some role in disambiguation (e.g., the word "ball" is more likely to be associated with the "toy" sense than with the "dance" sense).

There are some substantial differences, of course, between diagnostic inference and natural language processing. For example, the order of manifestations in diagnostic problem-solving is often insignificant to a plausible solution, but the order of words in a sentence is usually an important piece of information used in word sense disambiguation. Also, in diagnosis, if several hypotheses cannot be further discriminated, they may all be accepted as a tentative problem solution. In natural language processing a single coherent explanation is generally desired.

Despite these differences, the strong similarities between these two categories of problems suggest the possibility of applying parsimonious covering theory to solve certain natural language processing problems.

An exploratory study was undertaken to examine this issue [Dasigi & Reggia, 1989]. This work focused on an experimental prototype which automatically generates natural language interfaces for expert systems. The prototype is domain-independent in the same sense that a generic expert system shell is domain-independent. Given a knowledge base for a specific application, a vocabulary extractor extracts and indexes the linguistic information which it contains. In addition, an indexed domain-independent knowledge base that contains linguistic knowledge common to many domains is used. A natural language interface is generated for the specific application domain defined by the knowledge base using this knowledge plus a parsimonious covering inference mechanism.

Several modifications to PCT were introduced to handle some of the aspects of natural language processing that differ from diagnosis. By choosing representations in terms of descriptions for words that take the order of the words into account, the lack of sensitivity of PCT to word order is significantly improved. Unlike the entities in PCT, the different roles of words is taken into account in the hypothesis construction process. The syntactic (e.g., noun/adjective, subject/object) and semantic (word senses) aspects of covering are integrated in a mutually cooperative manner. In this system, a hypothesis is required, on the one hand, to parsimoniously cover words so that the syntactic descriptions in the linguistic knowledge base are satisfied, and on the other hand, to concurrently cover the words semantically with one or more domain-specific entities.

The prototype implementation suggests that it may be possible to combine effectively concurrent syntactic and semantic covering of natural language processing in the framework of an extended PCT. It remains to be seen if, ultimately, PCT can be extended to model abductive inference occurring in natural language processing in its full generality, namely, where the intentions and plans of the discourse participants and subjects are complex and play a significant role in understanding the discourse.

Conclusion

This paper has briefly considered some ways in which PCT, a formal model of abductive inference in diagnosis, can be related to abductive inference in general. It seems likely that the methods and principles used in PCT, extended to encompass a more general knowledge representation, could provide a fairly general theory of abductive inference. Such a theory would encompass not only diagnosis but also inference in natural language processing, legal inference, scientific discovery, etc. Producing a generalized PCT of this breadth would require a major research effort. We have indicated some of the significant generalization that would be needed elsewhere [Chu & Reggia, 1990; Peng & Reggia, 1990].

In its current form, PCT can still be useful to those attempting to develop other models/theories of abductive inference. As illustrated for the heuristics in TEC, PCT can serve as a useful "test case", suggesting potential limitations of a more general theory.

Acknowledgements. Supported by NIH Award NS-16332 and by NSF Award IST-8451430. Dr. Reggia is also with the Institute for Advanced Computer Studies and the Dept. of Neurology at the University of Maryland. Dr. Peng is also with the Institute for Software, Academia Sinica, Beijing, China.

References

Chu, B. and Reggia, J. (1990). Modeling Diagnosis at Multiple Levels of Abstraction, *Intl. J. Intell. Sys.*, in press.

Dasigi, V. and Reggia, J. (1989). Parsimonious Covering as a Method for Natural Language Interfaces to Expert Systems, *Artif. Intell. in Medicine*, 1, 49-60.

deKleer, J. and Williams, B. (1986). Reasoning About Multiple Faults. *Proc. 5th Nat. Conf. Artif. Intell.*, AAAI, 132-139.

Josephson, J., Chandrasekran, B., Smith, J., & Tanner, M. (1987). A Mechanism for Forming Composite Explanatory Hypotheses. *IEEE Trans. Sys., Man, and Cybern.*, 17, 445-454.

Peng, Y. and Reggia, J. (1987). A Probabilistic Causal Model for Diagnostic

Problem Solving. *IEEE Trans. Sys., Man and Cybern, 17*, 146-162 & 395-406.

Peng, Y. and Reggia, J. (1989). A Connectionist Model for Diagnostic Problem Solving. *IEEE Transactions on Systems, Man and Cybernetics*, 19, 285-298.

Peng, Y. and Reggia, J. (1990). *Abductive Inference Models for Diagnostic Problem Solving*, Springer-Verlag, in press.

Pople, H. (1973). On the Mechanization of Abductive Logic. *Proc. Internat. Joint Conf. Artificial Intelligence*,

Reggia, J., Nau, D., and Wang, P. (1983). Diagnostic Expert Systems Based on a Set Covering Model, *Int. J. Man-Machine Studies*, 19, 437-460.

Reggia, J., Nau, D., Wang, P., and Peng, Y. (1985). A Formal Model of Diagnostic Inference. *Infor. Sci., 37*, 227-285.

Reiter, R. (1987). A Theory of Diagnosis from First Principles. *Art. Intell.*, 32, 57-95.

Thagard, P. (1989). Explanatory Coherence. *Behav. and Brain Sci.*, 12, 435-502.

Wald, J., Farach M., Tagamets, M., and Reggia, J. (1989). Generating Plausible Diagnostic Hypotheses with Self-Processing Causal Networks, *J. Exp. and Theoret. AI*, 99-112.

Explanatory Coherence and Naturalistic Decision Making

Paul Thagard

Cognitive Science Laboratory, Princeton University
221 Nassau St., Princeton, NJ. 08542
paul@clarity.princeton.edu

This abstract briefly describes a theory of explanatory coherence and its implementation using a connectionist program called ECHO. I show how explanatory coherence considerations can play a large role in decision making in cases where decisions depend on evaluation of competing hypotheses. The abstract discusses the decision made in July 1988 by Captain Rogers of the USS Vincennes to shoot down what appeared to be an attacking aircraft. ECHO has been used to simulate the reasoning underlying this decision.

Explanatory Coherence

A theory of explanatory coherence (TEC) can be stated using the following seven principles (Thagard 1989). S is a system of propositions P, Q, and $P_1 \ldots P_n$. Local coherence is a relation between two propositions. I coin the term "incohere" to mean that two propositions are incoherent, which is stronger than saying that they do not cohere.

Principle 1. Symmetry.

(a) If P and Q cohere, then Q and P cohere.
(b) If P and Q incohere, then Q and P incohere.

Principle 2. Explanation.

If $P_1 \ldots P_m$ explain Q, then:

(a) For each P_i in $P_1 \ldots P_m$, P_i and Q cohere.

(b) For each P_i and P_j in $P_1 \ldots P_m$, P_i and P_j cohere.

(c) In (a) and (b) the degree of coherence is inversely proportional to the number of propositions $P_1 \ldots P_m$.

Principle 3. Analogy.

If P_1 explains Q_1, P_2 explains Q_2, P_1 is analogous to P_2, and Q_1 is analogous to Q_2, then P_1 and P_2 cohere, and Q_1 and Q_2 cohere.

Principle 4. Data Priority.

Propositions that describe the results of observation have a degree of acceptability on their own.

Principle 5. Contradiction.

If P contradicts Q, then P and Q incohere.

Principle 6. Acceptability.

(a) The acceptability of a proposition P in a system S depends on its coherence with the propositions in S.

(b) If many results of relevant experimental observations are unexplained, then the acceptability of a proposition P that explains only a few of them is reduced.

Principle 7. System coherence.

The global explanatory coherence of a system S of propositions is a function of the pairwise local coherence of those propositions.

Principle 2, Explanation, covers cases where hypotheses explain evidence or are themselves explained by higher level hypotheses. Clauses 2(a) and 2(b) state that hypotheses that explain a proposition cohere with that proposition and with each other. Clause 2(c) is a simplicity principle, suggesting that the greater the number of hypotheses needed to explain a proposition, the less they cohere with it and with one another. Principle 3, Analogy, says that similar hypotheses that explain similar pieces of evidence cohere. The fourth principle, Evidence, is straightforward. Principle 5, Contradiction, marks competing hypotheses as incoherent with each other if they

are contradictory. The last two principles state that the previous five principles establishing local relations of explanatory coherence are all that is needed to determine the overall coherence of a set of propositions and the acceptability of particular propositions. These contentions have been put to the test by development of a computer program that allows simulation of judgments of explanatory coherence.

ECHO

Connectionist networks consist of units, roughly analogous to neurons, that are connected by excitatory and inhibitory links (Rumelhart and McClelland, 1986). ECHO is a Common LISP program that constructs networks for evaluating the explanatory coherence of sets of propositions. Propositions that cohere are represented by units connected by excitatory links, while ones that incohere have units connected by inhibitory links. For input, ECHO is given formulas describing the explanatory relations of propositions. If two hypotheses H1 and H2 together explain a piece of evidence E1, ECHO is given the LISP input:

(EXPLAIN '(H1 H2) E1).

In accord with the second principle of explanatory coherence, ECHO then sets up symmetric excitatory links between units representing H1 and E1, H2 and E1, and H1 and H2. If H1 and H3 are contradictory, ECHO gets the input:

(CONTRADICT 'H1 'H3).

This sets up a symmetric inhibitory link between H1 and H3. That E1 and E2 are to be treated as pieces of evidence is represented by the input:

(DATA '(E1 E2))

In accord with principle 4, Evidence, links are then set up from a special evidence unit to E1 and E2.

Connectionist networks make decisions by repeatedly updating the activation of units in parallel until the whole network settles into a stable state in which the activation of each unit has reached asymptote. ECHO adjusts the activation of a unit u_j by considering all the units to which it is linked. An excitatory link with an active unit will increase the activation of u_j, while an inhibitory link with a unit with positive activation will decrease it. Activation of units starts at 0 and is allowed to range between 1 and -1. Repeated adjustments of activations results in a stable state where some units end up with high activation and others with activation below 0. Equations and algorithms are fully presented elsewhere (Thagard 1989). Parallel constraint-satisfaction techniques similar to ECHO's have also proven useful for investigating analogy (Holyoak and Thagard 1989; Thagard, Cohen, and Holyoak 1989; Thagard, Holyoak, Nelson, and Gochfeld, in press).

ECHO has been used to simulate many cases of scientific and legal reasoning (Thagard 1989, forthcoming; Thagard and Nowak 1988, forthcoming; Ranney and Thagard 1988; Nowak and Thagard forthcoming-a, forthcoming-b). I will now briefly describe its application to naturalistic decision making.

Naturalistic Decision Making

Decision making can sometimes straightforwardly take place using an assessment of possible actions with respect to probabilities and utilities of results of the actions. Often, however, it is necessary to form and evaluate hypotheses concerning the nature of the situation. For example, a fire chief may need to infer the source and nature of a fire before deciding how best to fight it. Although in AI it is becoming common to refer to both the formation and the evaluation of explanatory hypotheses as *abduction*, I shall follow the use of its inventor C.S. Peirce and reserve the term for hypothesis formation only (Thagard 1988). TEC and ECHO are concerned, not with abduction in this narrow sense, but with hypothesis *evaluation*.

Judges and juries are frequently called upon to evaluate explanatory hypotheses in criminal trials, asking, for example, whether the proposition that the accused murdered the deceased is the best explanation of the death and other evidence. But inference to the best explanation in such cases is not just a matter of considering what hypothesis explains the most evidence, since it is standard in trials to consider a *motive* that could explain why the murder was committed. The acceptability of a hypothesis increases on the basis of there being explanations of it, as well as on the basis of what it explains. Everyday decisions that involve other people often involve explanatory inferences concerning their beliefs, desires, and intentions. In adversarial situations such as competitive games, business, diplomacy, and war, it is often necessary to infer the plans of the adversary. Plans can sometimes be inferred as part of the best explanation of what the adversary has done so far.

Let us now look in more detail at an actual case of a decision that is naturally understood in terms of explanatory coherence. On July 3, 1988, the USS Vincennes was involved in a battle with Iranian gunboats in the Persian Gulf. A plane that had taken off from Iran was observed to be flying toward the Vincennes. On the basis of the information provided to him by his officers, Captain Rogers of the Vincennes concluded that the plane was an attacking Iranian F-14 and shot it down. Unfortunately, the plane turned out to be a commercial flight of Iran Air 655. Nevertheless, the official investigation (Fogarty 1988) concluded that Rogers acted in a prudent manner. An ECHO-analysis of the information available to Rogers

supports that conclusion.

Rogers' decision to fire a missile at the plane depended on evaluation of competing hypotheses concerning its nature and intentions. The hypothesis that it was a commercial flight was considered and rejected in favor of the hypotheses that the plane was an F-14 and that it was attacking. Captain Rogers recalled numerous "indicators" used in declaring the plane hostile and deciding to engage (Fogarty 1988, p. 40). From the perspective of TEC, the F-14 hypotheses were more coherent than the alternatives for several reasons. First, they explained why the plane did not respond to verbal warnings, was not following commercial air corridors, was veering toward the Vincennes, and was reported to be descending. (This report turned out to be erroneous.) Second, the commercial-airline hypotheses predicted (explained) the negation of this evidence. Finally, the F-14 attack could be explained by hostile Iranian intentions for which there was ample evidence.

Here is the actual input given to ECHO. Note that the quoted propositions are for information only: unlike a program that would be capable of forming the hypotheses and generating hypotheses about what explains what, ECHO does not use the content of the propositions. For ease of cross-reference, I have numbered propositions in correspondence to the list in the Fogarty report (p. 40), although a few of the pieces of evidence do not appear relevant to an assessment of explanatory coherence.

EVIDENCE:
(proposition 'E0 "Gunboats were attacking the Vincennes.")
(proposition 'E1 "F-14's had recently been moved to Bandar Abbas.")
(proposition 'E2 "Iranian fighters had flown coincident with surface engagement on 18 April 1988.")
(proposition 'E3 "The aircraft was not responding to verbal warnings over IAD or MAD.")
(proposition 'E4 "There had been warnings of an increased threat over the July 4 weekend.")
(proposition 'E5 "There had been a recent Iraqui victory.")
(proposition 'E6 "The aircraft was not following the air corridor in the same manner as other commercial aircraft had been seen consistently to behave.")
(proposition 'NE6 "The aircraft was flying in the commercial air corridor.")
(proposition 'E7 "The aircraft was flying at a reported altitude which was lower than COMAIR was observed to fly in the past.")
(proposition 'NE7 "The aircraft flew at COMAIR's usual altitude.")
(proposition 'E8 "Track 4131 was reported to be increasing in speed.")
(proposition 'E9 "Track 4131 was reported to be decreasing in altitude.")
(proposition 'NE9 "Track 4131 was reported to be increasing in altitude.")
(proposition 'E10 "Track 4131 was CBDR to USS Vincennes and USS Montgomery.")
(proposition 'E11 "Track 4131 was reported by USS VINCENNES' personnel squawking Mode II-1100 which correlates with an F-14.")
(proposition 'E12 "No ESM was reflected from track 4131.")
(proposition 'E13 "F-14s have an air-to-surface capability with Maverick and modified Eagle missiles.")
(proposition 'E14 "The aircraft appeared to be maneuvering into attack position; it veered toward the USS Montomery.")
(proposition 'E15 "deleted in published report")
(proposition 'E16 "Visual identification of the aircraft was not feasible.")
(data '(E0 E1 E2 E3 E4 E5 E6 E7 E8 E9 E10 E11 E12 E13 E14 E15 E16))

HYPOTHESES:
(proposition 'A1 "Iran is intending to mount an attack.")
(proposition 'A2 "The plane is an F-14.")
(proposition 'A3 "The plane intends to attack.")
(proposition 'A4 "The F-14 is flying 'cold-nose'.")
(proposition 'C1 "The plane is a commercial airliner.")
(proposition 'C2 "The plane is taking off.")

EXPLANATIONS:
(explain '(A1) 'E0)
(explain '(A1) 'E1)
(explain '(A1) 'E4)
(explain '(A1) 'A3)
(explain '(A1) 'A2)
(explain '(A2 A3) 'E3)
(explain '(E5) 'A1)
(explain '(A2) 'E6)
(explain '(C1) 'NE6)
(explain '(A2) 'E7)
(explain '(C1) 'NE7)
(explain '(A2 A3) 'E8)
(explain '(C1 C2) 'E8)
(explain '(A2 A3) 'E9)
(explain '(C2) 'NE9)
(explain '(A3) 'E10)
(explain '(A2) 'E11)
(explain '(A2 A4) 'E12)
(explain '(C1) 'E12)
(explain '(A3) 'E14)

CONTRADICTIONS:
(contradict 'E6 'NE6)
(contradict 'E7 'NE7)
(contradict 'E9 'NE9)
(contradict 'A2 'C1)

Missing from this analysis is the possible impact of analogy to previous incidents: the Fogarty report mentions the Stark incident that involved the sinking of an American ship in 1987. One can easily imagine Rogers reasoning that just as the Stark should have explained the behavior of an approaching Iranian plane in terms of its hostile intentions, so should he give an analogous explanation in the current case. TEC (principle 3) and ECHO can naturally model the impact that such analogies can have.

Figure 1 displays the network that ECHO creates using this input, showing the excitatory and inhibitory links. Figure 2 graphs the activation of the units over the 60 cycles that it takes the network of settle, showing that the units A2 and A3 concerning an attack by an F-14 become activated while C1 and C2 concerning a commercial airliner are deactivated.

Because the hypotheses of an attacking F-14 was more coherent with the available information than the commercial-airline hypothesis, and because F-14s were known to be capable of severely damaging the Vincennes, Captain Rogers shot the plane down. The fact that a tragic mistake was made does not undermine the fact that the evidence pointed strongly toward an F-14. The Fogarty report found fault with the ship's Tactical Information Coordinator and Anti-Aircraft Warfare officer for providing Rogers with the erroneous information that the plane was descending rather than ascending, but this was only one factor in making the F-14 hypothesis more plausible.

Thus the decision on the USS Vincennes can be understood in terms of the theory of explanatory coherence. Future research will perform additional analyses and simulations to determine the applicability of explanatory-coherence factors to decision making in legal, military, and everyday contexts.

Acknowledgements. This research is funded by a contract from the Basic Research Office of the Army Research Institute for Research in the Behavioral and Social Sciences, and by a grant from the McDonnell Foundation to the Princeton University Human Information Processing Group.

References

Fogarty, W. (1988). Formal investigation into the circumstances surrounding the downing of Iran Air Flight 655 on 3 July 1988. Washington: Department of Defense.

Holyoak, K. and Thagard, P. (1989) Analogical mapping by constraint satisfaction. *Cognitive Science 13*, 295-355.

Nowak, G., and Thagard, P., (forthcoming). Copernicus, Ptolemy, and explanatory coherence. For a volume of *Minnesota Studies in the Philosophy of Science*.

Nowak, G., and Thagard, P., (forthcoming). Newton, Descartes, and explanatory coherence. Manuscript in preparation.

Ranney, M., and Thagard, P. (1988). Explanatory coherence and belief revision in naive physics. *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum. 426-432.

Rumelhart, D. E., McClelland, J. R., and the PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*, 2 volumes. Cambridge, Mass.: MIT Press.

Thagard, P. (1988). *Computational Philosophy of Science*. Cambridge, Mass.: MIT Press/Bradford Books.

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences, 12*, 435-467.

Thagard, P. (forthcoming) The dinosaur debate: Explanatory coherence and the problem of competing hypotheses. To appear in J. Pollock and R. Cummins, eds., *Philosophy and AI: Essays at the Interface*.

Thagard, P., Cohen, D., and Holyoak, K., (1989) Chemical analogies: Two kinds of explanation. *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*. San Mateo: Morgan Kaufmann, 819-824.

Thagard, P., Holyoak, K., Nelson, G., and Gochfeld, D. (in press). Analog retrieval by constraint satisfaction. *Artificial Intelligence*.

Thagard, P., and Nowak, G. (1988). The explanatory coherence of continental drift. In A. Fine and J. Leplin (Eds.), *PSA 1988*, vol. 1. East Lansing, Mich.: Philosophy of Science Association, 118-126.

Thagard, P., and Nowak, G. (forthcoming). The conceptual structure of the geological revolution. In J. Shrager and P. Langley, eds., *Computational Models of Discovery and Theory Formation*. Hillsdale, NJ: Erlbaum.

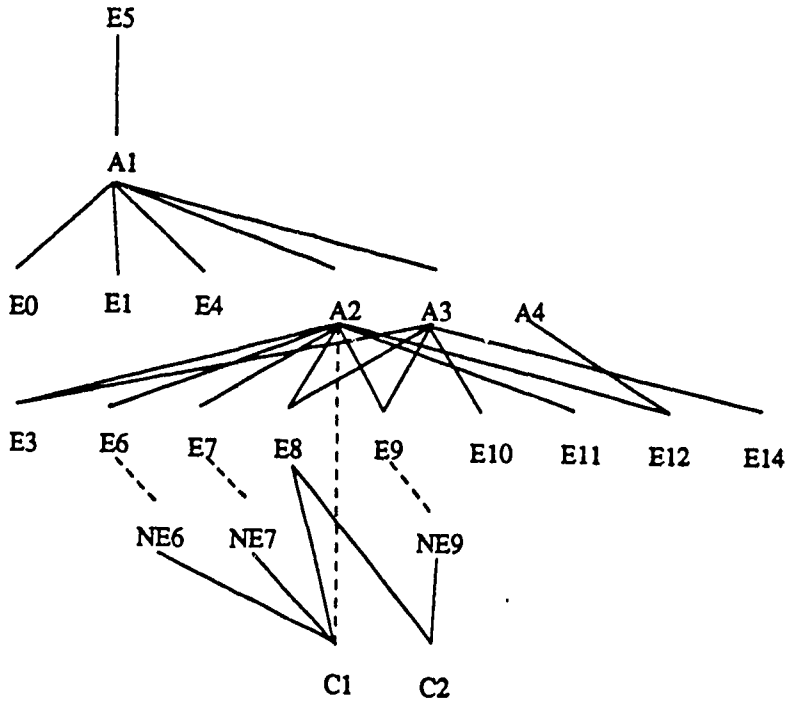


Figure 1. Network created by ECHO using Vincennes input. Solid lines indicate excitatory links, while dotted lines indicate inhibitory links. A1-A3 represent units hypothesizing an attack. C1-C2 represent units concerning a commercial airliner.

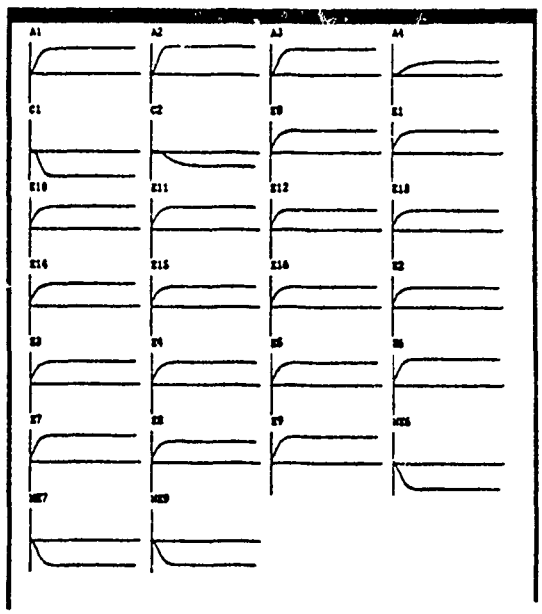


Figure 2. Activation history of units in Vincennes simulation. Each graph shows the activation of a unit over 110 cycles of updating, on a scale of -1 to 1, with the horizontal line indicating the initial activation of 0. The network has settled by 60 cycles.

Abduction in Model Generative Reasoning

Roger T. Hartley and Michael J. Coombs

Computing Research Laboratory and
Computer Science Department
New Mexico State University
Las Cruces, NM 88003

Model Generative Reasoning

At CRL, we have developed the Model Generative Reasoning (MGR) architecture in order to study the relationship between structure and semantics in coping with brittleness (Coombs and Hartley, 1987; 1988). Formally, MGR is related to the generalized set covering (GSC) model of abductive problem solving, where, given knowledge of a set of observations (facts), the task is to find the best explanatory hypothesis in terms of the most parsimonious "cover" of facts by hypotheses. However, whereas GSC deals with atomic explanatory hypotheses and pre-defined relevance relations between hypotheses and facts, it is necessary in a noisy or novel task environment to: (i) create new hypotheses from conceptual fragments, and (ii) identify problem facts as some subset of the set of available observations.

Hypotheses, explanation and abduction

All problem solvers generate hypotheses, and in general we can classify all such mechanisms as abductive. However, the use to which these subsequent hypotheses are put separates what we might called logical abduction from the more pragmatic use of the term in scientific reasoning (Peirce, 57).

Many authors have pointed out that abduction is an unsound logical inference from consequent to antecedent, as in:

$$\frac{B \qquad A \rightarrow B}{A}$$

Such a method is used extensively in backward-chaining expert systems and is the basis of Prolog's proof technique. Most of these, however, set up the A's as intermediate goals to be carried on further in the method i.e. they are not asserted as true, or as possibly true. Another way to look at such a mechanism is that if B is true, and A $\rightarrow$ B, then A ***explains*** B. This is the basis of the methods described by Levesque (Levesque, 89) and Poole (Poole, 89). Explanations are hypothetical structures generated to fit some set of observations, (not just one, as the above simplification implies) and have only possible status in the system. However, if they were true, then the consequent would follow naturally by deduction. Finding an explanation in a logical system then amounts to finding an expression, that if it were true would imply the input (the axioms). In order to find out how this works, we need to analyze the use of rules (logical implications) in such systems.

There are four main uses of a rule:

- as a selectional constraint on types, e.g. all U's are V's:

$$\forall x\, U(x) \rightarrow V(x) \qquad (1)$$

- as an Aristotelian definition, e.g. if something has properties A, B, C etc. then it is a V

$$\forall x\, A(x) \wedge B(x) \wedge C(x) \cdots \rightarrow V(x) \qquad (2)$$

- as a contingent, or schematic definition, e.g. if something is a V, then it has properties A, B, C etc.

$$\forall x\, V(x) \rightarrow A(x) \wedge B(x) \wedge C(x) \cdots \qquad (3)$$

- to express causality, e.g. if P, Q, R all happen,

then X, Y, Z will happen as a direct consequence

$$P \wedge Q \wedge R.... \rightarrow X \wedge Y \wedge Z \cdots \quad (4)$$

To illustrate these types of rule, consider the following abductive inferences:

$$\frac{car(a) \qquad \forall x\ ford(X) \rightarrow car(x)}{possible(ford(a))} \quad (5)$$

In other words, if a car a is observed, then it is possibly a ford. Of course, this is not an explanation of *why a* is a car, but it does shed a little more light on the subject. There may well be other types of car (chevrolet, subaru etc.) and these would be equally likely inferences. The question of the goodness of an explanation is one for the pragmatics of abduction. Poole has pointed out, however, (Poole, op. cit.) that there are several possible accounts of what constitutes the best explanation.

$$\frac{car(a) \qquad \forall x\ hasWheels(x) \wedge hasEngine(x) \wedge hasDriversSeat(x) \rightarrow car(x)}{possible(wheels(a) \wedge hasEngine(a) \wedge hasDriversSeat(a))} \quad (6)$$

This abductive inference stands a little better as an explanation; it at least shows why a is a car, based on the limited knowledge to hand about cars. Note that if both of the above rules were available and a criterion of goodness was expressed as the simplest explanation is the best, then the first would be preferred over the second.

$$\frac{at(a,b) \qquad \forall xyz\ person(x) \wedge car(y) \wedge drive(x,y) \wedge location(z) \rightarrow at(x,z) \wedge at(y,z)}{possible(\exists y\ person(a) \wedge car(y) \wedge drive(a,y) \wedge location(b))} \quad (7)$$

Here we infer an existential result (as pointed out by Poole) because we have no evidence about the car if we assume a is a person. The other possible inference is:

$$possible(\exists y\ person(y) \wedge car(a) \wedge drive(y,a) \wedge location(b)) \quad (8)$$

It is also possible that the observations are existentially quantified. e.g. with the fact:

$$\exists xy\ at(a,x) \wedge drive(y,b) \quad (9)$$

This, with the same rule as in 3, gives the inference:

$$possible(\exists x\ person(x) \wedge car(b) \wedge drive(a,b) \wedge location(x)) \quad (10)$$

Finally an example that gets closer, we believe to the reason why abduction is important. If we have the facts:

$$at(a,b)\ and \quad (11)$$

$$hasEngine(c) \quad (12)$$

which seem to be unconnected, the job of abduction is to *glue* them together in a single hypothesis. The inference we might look for, and one that is clearly an explanation of why these two pieces of data are observed together is:

$$person(a) \wedge car(c) \wedge drive(a,c) \wedge location(b) \wedge at(c,b) \quad (13)$$

This hypothesis glues the facts together, through the definition of a car and the driving rule.

An operator for abduction: specialize

We will describe an single operator, *specialize*, which mechanizes the process of abduction illustrated in the above examples. It is composed of two more primitive operators, *cover* and *join* that operate on conceptual graphs (Sowa, 84). This leads to the slogan:

Abduction = cover + join

Conceptual graphs are connected, directed, bipartite graphs where the nodes are labeled with either a concept type or a relation name. There are restrictions on the edges, however, that are used to preserve semantic coherence (Sowa calls this *canonicality*). A relation node may have only one ingoing edge, but any number of outgoing edges. A concept node may have any number of edges, in or out.

The functionality of specialize is:

$$specialize : 2^{\mathcal{F}} \times 2^{\mathcal{D}} \rightarrow 2^{\mathcal{H}} \quad (14)$$

where $\mathcal{F}$ is a set of input graphs, $\mathcal{D}$ is a set of *definitions* (conforming to the rule types 2, 3 and 4 above) and $\mathcal{H}$ is the resultant set of hypotheses produced by cover and join.

The operator cover

It is the job of cover to choose an appropriate subset of a set of stored graphs $\mathcal{D}$, that cover all of the concepts in a given subset of graphs taken from a set $\mathcal{F}$. If the conceptual content of a graph g is given by $C(g)$ and the maximal common subtype of two concepts c_1 and c_2 is given by $M_b(c_1, c_2)$ then the functionality of cover is given by:

$$cover : \mathcal{F} \times 2^{\mathcal{D}} \rightarrow 2^{\mathcal{D}} \qquad (15)$$
$$where\ for\ f \in \mathcal{F},\ \forall c \in C(f) \exists c_d, d \mid$$
$$c_d \in C(d)\ for\ some\ d \in \mathcal{D}$$
$$and\ M_b(c, c_d)\ exists$$

In other words, every concept in f must have at least one concept in the set of graphs $\mathcal{D}_C$, where their maximum common subtype exists i.e. is not bottom. There are problems with graphs containing duplicate labels, but these can be solved by ensuring that there are sufficient quantities of covering concepts from graphs in $\mathcal{D}$ for the concepts in f.

The choice of an appropriate subset, since there can be many which satisfy the above condition is a matter for the pragmatics of the problem. The Maryland group (Nau and Reggia, 86) have used this idea of set covering (as have many others) in their diagnostic work, but deal with expressions at the propositional level rather than at the object level as we do here. They point out that although a parsimonious cover may be appropriate when simplicity is called for (cf. Occam's razor) there be cases when less than parsimonious cover er, or simply better as an explanation.

arsimonious cover may be produced by minimizing the boolean expression:

$$\bigwedge_c \bigvee_i d_i, where\ c \in C(f), C(d_i) \qquad (16)$$

The operator join

Cover just produces an appropriate subset of $\mathcal{D}$. The job of producing an explanatory hypothesis is left to the binary operation join (actually *maximal* join). As an operation on single concept nodes, join merges two graphs at a single point where both graphs contain the same concept label. *Maximal* join (we will usually refer to this as just *join*) will not only allow restrictions in that a concept label can be replaced by a label of any subtype but also will merge the two graphs on the maximum number of nodes (see Sowa, op cit). An example

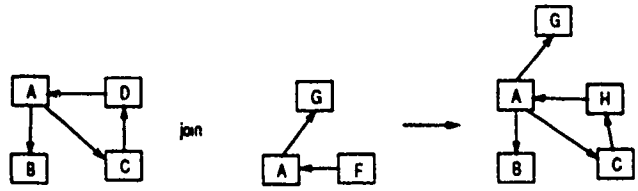


Figure 1: An example of maximal join ($M_b(D, F) = H$).

is given in Figure 1. The functionality of join is:

$$maximal\ join : \mathcal{G} \times \mathcal{G} \rightarrow 2^{\mathcal{G}} \qquad (17)$$

There can be more than one maximal join, hence the powerset notation on the set of all graphs $\mathcal{G}$. Join is a binary operation but multiple graphs can be joined by composing it with itself. Unfortunately, there is good reason to believe that join is not commutative when semantic considerations come into play (Pfeiffer and Hartley, 89), but for now we will assume there is no problem.

Since restrictions are allowed, it is clear that two nodes are joinable as part of a maximal join operation if they contain types that have a maximal common subtype. So PET can be restricted to DOG, and so can MAMMAL. Thus nodes containing PET and MAMMAL join to produce DOG. If two concepts have only $\perp$ (bottom) as their common subtype, then the maximal common subtype is not considered to exist. The reason why the same constraint was placed on the operator cover was so to ensure that the covers returned are maximally joinable i.e. that the fact graph f is joinable to the graphs that cover it. That this is an abductive inference in the logical sense may be seen from the following equivalent presentation:

$$\frac{PET(a) \quad MAMMAL(a) \quad \forall x DOG(x) \rightarrow PET(x) \quad \forall x DOG(x) \rightarrow MAMMAL(x)}{DOG(a)}$$

If we now look at the last car example above, the facts might be represented as in Figure 2, and the covering graphs in Figure 3. These graphs add information that the logical representation leaves out, however these are mandated by the need to

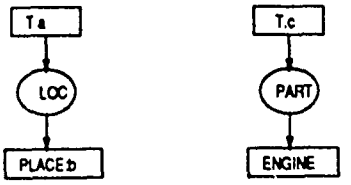


Figure 2: The 'driving' facts.

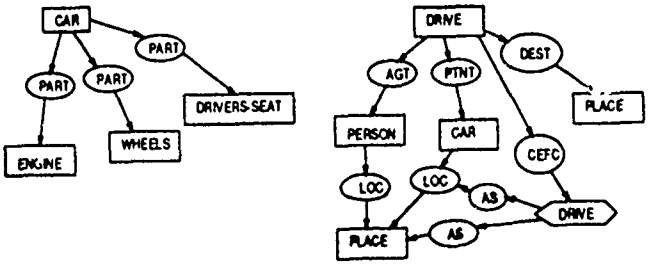


Figure 3: The covering graphs for Fig. 3.

form canonical graphs. The equivalent in a logic would be a full intensional logic with type restrictions on the place-holders, but the graphs have not prejudiced the argument that the appropriate hypothesis is obtained by joining all four graphs together, as shown in Figure 4. The major addition to the driving graph is the actor node in the diamond-ended box. In a extension to conceptual graphs (Hartley, forthcoming), these nodes can express the temporal relationships between states and events in order to represent procedures qualitatively. However, these extra nodes play no part in cover or join, and will not be discussed further. The join will only occur if the following type relationships hold:

$$M_b(CAR, PHYS-OBJ) = CAR$$
$$M_b(CAR, TRANSPORT) = CAR$$

It should be noted that a person, an engine and a drivers-seat are all physical-objects, in addition to the car. These relationships potentially give alternative joins. Thus instead of placing the person a at the place b, the join could place any of the

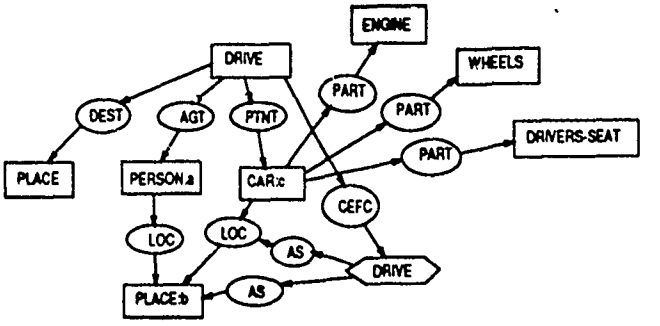


Figure 4: The maximal join of Figs, 3 and 4.

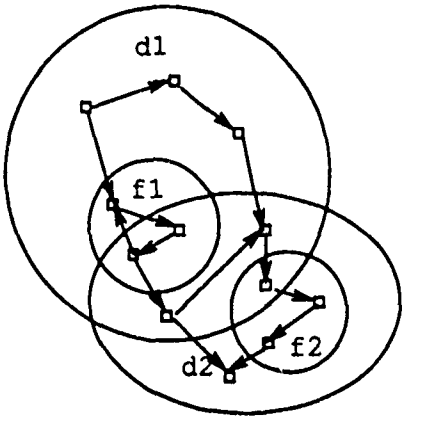


Figure 5: Specialize in a set diagram form.

other objects, for instance the engine, at b. This sort of thing may produce a violation of canonicality (e.g. giving a pipe three ends instead of two ends and a middle), but may also be prevented by knowledge of the *conformity* of individuals to types. Again, the FOPC form does not contain this information, but a may conform to PERSON, but not to ENGINE.

In essence, therefore, the resultant graphs produced by join can be seen as abductive inferences from the facts and definitions, causal or Aristotelian that cover them. The result is hypothetical in nature because the maximal common subtype restriction of two types leads to the same unsound inference rule that an logical abductive rule makes. Additionally, however, constraints stemming from canonicality and conformity increase the likelihood of the inference. Figure 5 contains a more intuitive Venn-like diagram of specialize where each enclosed region contains at least one concept node.

References and Bibliography

References

[1] Coombs, M. J. and R. T. Hartley (1987) The MGR algorithm and its application to the generation of explanations for novel events. *International Journal of Man-Machine Studies* 27: 679-708.

[2] Coombs, M. J. and R. T. Hartley (1988) Explaining novel events in process control through model generative reasoning. *International Journal of Expert Systems* 1: 89-109.

[3] Coombs, M. J. and R. T. Hartley (1988) Design of a software environment for tactical situation development. *Proceedings of the US Army Symposium on Artificial Intelligence Research for Exploitation of the Battlefield Environment*, El Paso, November 1988.

[4] Hartley, R.T. and Coombs, M.J. (1989). Reasoning with graph operations. Proc. Workshop on the formal foundations of semantic networks. Catalina Island. Also available as MCCS-89-166, Computing Reseatch Lb. New Mexico State University.

[5] Fields, C. A., M. J. Coombs and R. T. Hartley (1988). MGR: An architecture for problem solving in unstructured task environments. *Proceedings of the Third International Symposium on Methodologies for Intelligent Systems*, Elsevier, Amsterdam, 44-49.

[6] Fields, C. A., M. J. Coombs, E. S. Dietrich, and R. T. Hartley (1988b) Incorporating dynamic control into the Model Generative Reasoning system. *Proc. ECAI-88*, pp. 439-441.

[7] Hartley, R. T. and M. J. Coombs (1988) Conceptual programming: Foundations of problem solving. In: J. Sowa, N. Foo, and P. Rao (Eds) *Conceptual Graphs for Knowledge Systems*. Reading, MA: Addison-Wesley (in press). Also available as MCCS-88-129, Computing Research Lab., New Mexico State University.

[8] Levesque, H. (1989) A knowledge-level account of abduction. *Proc. Eleventh International Joint Conference on Artificial Intelligence*, Detroit, MI. pp. 1061-1073.

[9] Nau, D. and J. Reggia (1986) Relationships between deductive and abductive inference in knowledge-based diagnostic problem solving, in L. Kerschberg (Ed.), *Expert Database Systems: Proceedings of the First International Workshop*. New York: Benjamin Cummings.

[10] Peirce, C.S. (1957). *Essays in the Philosophy of Science*. New York: Bobbs-Merrill.

[11] Pfeiffer, H.D. and Hartley, R.T. (1989) Semantic additions to conceptual programming. Conceptual Graphs Workshop, IJCAI89.

[12] Poole, D. (1989) Explanation and prediction: an architecture for default and abductive reasoning. *Computational Intelligence* (5), pp. 97-110.

[13] Sowa, J.F. (1984). *Conceptual Structures*. Reading, MA: Addison Wesley.

Abduction as Similarity-Driven Explanation

Brian Falkenhainer
Xerox Palo Alto Research Center
3333 Coyote Hill Road, Palo Alto, CA 94304

1 Introduction

Deduction, abduction, and analogy are processes whose differences are normally reflected by distinct computational mechanisms. In this paper, I suggest that procedural separations between these processes are superfluous for the purpose of constructing plausible explanations of a given phenomenon. A single mechanism that proposes explanations of phenomena by their similarity to understood phenomena is sufficient, providing smoother adaptability to unanticipated or underspecified events and enabling transfer of knowledge from one domain to another. This *similarity-driven* view of explanation also lets one extend or revise imperfect theories when they fail to produce an explanation. In this approach, one provides a deductive explanation if possible, and extends or revises the underlying theory when necessary to make explanation possible. Rather than being produced by separate processes, distinctions between the different explanation types result from the preferential ordering imposed when competing hypotheses are evaluated.

The plausibility of this conjecture is demonstrated by PHINEAS, a program that uses a single similarity-driven explanation mechanism to focus its search for explanations using its existing knowledge and to develop novel theories when its existing knowledge is insufficient.

This paper begins with a discussion of the relationship between abductive explanation and analogy, suggesting that they share a common core: the search for explanatory similarity. It then briefly describes PHINEAS and outlines some of the examples used to test its behavior.

2 Similarity-Driven Explanation

Abduction is traditionally characterized as using a fixed set of background theories. Assumptions needed to fill gaps due to incomplete knowledge of the situation are limited to ground atomic sentences (i.e., no new or revised rules are considered), as in

given CAUSE($\mathcal{A}, \mathcal{C}$), $\mathcal{C}$ *infer* $\mathcal{A}$

These systems suffer from the *adaptability problem*: they are unable to revise or extend an imperfect domain theory to make conjectures about unanticipated events, and unable to apply knowledge of one domain to the understanding of another.

On the other hand, theory formation typically involves making assumptions about both the situation and the incompleteness or incorrectness of current theories. It includes inferences of the form

given CAUSE($\mathcal{A}, \mathcal{C}$) $\wedge$ $\mathcal{A} \Rightarrow \mathcal{C}$, $\mathcal{A}$, $\mathcal{C}$
infer CAUSE($\mathcal{A}, \mathcal{C}$)

Theory formation must face the problem of generating theory-revising hypotheses and establishing a preference among a possibly infinite set of hypotheses.

To address these problems, we note the strong commonalities between traditional abduction and analogy, and develop a model that encompasses both. For abduction, this unified model provides the power to extend the underlying domain theory when needed. For theory formation, it enables existing knowledge, possibly of other domains, to influence hypothesis generation and evaluation, thus taking into account knowledge of the way things normally behave in the world and the way theories about those behaviors are normally expressed. This view of explanation is based on the conjecture that search for similarity between the situation being explained and some understood phenomenon suffices as the central process model for explanation tasks.

In support of this view, consider the explanation scenarios summarized below:

Deduction scenario: Given phenomenon $\mathcal{P}$, where $\mathcal{P}$ represents a set of observables, a complete explanation of $\mathcal{P}$ deductively follows from existing knowledge. The only open question is whether it is *the* explanation, as there may be others. For example, suppose fluid flow is observed and all of the preconditions for fluid flow are known to hold (e.g., the source pressure is greater than the destination pressure, the fluid path is open, etc.). Then a fluid flow explanation directly follows. Given the observed behavior and the existing preconditions, we could say that the situation is *literally similar* (Gentner, 1983) to liquid flow.

Assumption scenario: Phenomenon $\mathcal{P}$ is given, where $\mathcal{P}$ represents a set of observables. No explanation can be found using current knowledge because the status of some requisite facts is unknown. However, a complete explanation follows from the union of existing knowledge and a consistent set of assumptions about the missing facts. For example, if one observes liquid flow but does not know if the fluid path valve is open or closed, one can assume the valve is open if there is no evidence to the contrary.

Generalization scenario: Phenomenon $\mathcal{P}$ is given, where $\mathcal{P}$ represents a set of observables. Existing knowledge indicates that candidate explanation $\mathcal{E}$ cannot apply because condition C_1 is known to be false in the current situation. However, $\mathcal{E}$ does follow if condition C_1 is replaced by the next most general relation, since C_1's sibling is true in the current situation. This is a standard knowledge-base refinement scenario.

Analogy scenario: Phenomenon $\mathcal{P}$ is given, where $\mathcal{P}$ represents a set of observables. No candidate explanation $\mathcal{E}$ is available directly, but $\mathcal{E}_b$ is available if a series of analogical assumptions are made, that is, if the situation explained by $\mathcal{E}_b$ is assumed analogous to the current situation. For example, if heat flow is observed, but little is known about heat phenomena, an explanation may be constructed by analogy to liquid flow.

Each scenario requires the interpretation-construction task: retrieve from memory explanatory hypotheses that match the current situation. Each also requires the interpretation-selection task: select from a set of candidate hypotheses the one that is most probable, plausible, or coherent. Importantly, each scenario represents the same process when viewed as different forms of similarity to an existing theory:

- *Deduction scenario*: complete match of identical features
- *Assumption scenario*: partial match of identical features
- *Generalization scenario*: matches between features having a close generalization
- *Analogy scenario*: a range of matches between different features and relations

A system based on this view would offer the best explanation available, ranging from application of an existing theory to distant analogy. It assumes that all interpretation-construction tasks may be characterized as the search for maximal, explanatory similarity between the situation being explained and some previously explained scenario. It suggests using a single computational architecture for explanation processes. Distinctions between explanation types only influence the weighing of evidence and the decision as to whether a new conjecture represents a revision of existing knowledge or a new separate body of knowledge.

3 The PHINEAS System

The similarity-driven model of explanation discussed in the previous section is illustrated by **PHINEAS**, a program that offers qualitative explanations of time-varying physical behaviors. The system uses remindings of similar experiences to suggest plausible hypotheses, and uses qualitative simulation to analyze the consistency and adequacy of these hypotheses.

PHINEAS uses three sources of knowledge during its reasoning process. First, its uses an initial domain theory consisting of a collection of qualitative theories about physical processes (e.g., liquid flow), entities (e.g., fluid paths), and general physical principles (e.g., mechanical coupling). This qualitative knowledge is represented using the language of Forbus' (1984) QP theory. Second, when comparing a new observation to prior experience, **PHINEAS** consults a library of previously observed phenomena (i.e., structure and behavior descriptions). The final source of **PHINEAS**' information is the observation targeted for explanation, which includes the original scenario description (e.g., `Open(beaker)`), the behavior across time (e.g., `Decreasing[Amount-of(alcohol)]`), and behavioral abstractions that apply to the observation (e.g., **asymptotic**).

In response to a given observation, **PHINEAS** attempts to produce an explanatory "theory" and the envisioned behaviors it predicts. A theory consists of a set of process descriptions, entity descriptions, and atomic facts. The process and entity descriptions may be elements of the existing domain theory or new postulated theories. The system makes this distinction during hypothesis evaluation. The atomic facts are assumptions about the scenario required to complete the explanation.

PHINEAS operates in four stages (see Falkenhainer, 1988 for more details):

Access. A new observation triggers a search in memory for understood phenomena that exhibit analogous behavior. This retrieval process involves two stages. First, behavioral abstractions of the observed situation are used to provide indices to a potentially relevant subset of memory. Second, each phenomenon in this sub-

set is inspected more carefully by matching its detailed structural and behavioral description to the current situation. This partial mapping provides an indication of what objects and quantities correspond by virtue of their behavioral similarity, and will serve as an important source of constraint during the mapping process. The match also indicates where the phenomena correspond and thus what portion of the base analogue's behavior should be considered relevant.

Mapping & Transfer. The objective of the second stage is to generate an initial hypothesis about the current observation. This has two components. First, the models used to explain analogous aspects of the recalled experience are retrieved and analogically mapped into the current domain. This mapping is guided by the initial correspondences found during access. Second, any unknown entities and properties in the hypothesis must be inferred from the domain theory or their existence must be postulated. The model of mapping used in this work is called *contextual structure-mapping* (Falkenhainer, 1988), a knowledge-intensive adaptation of Gentner's (1983) structure-mapping theory of analogy. Comparisons are performed by SME (Falkenhainer, Forbus, & Gentner, 1989).

Qualitative simulation. The predictions of a proposed model are compared against the observed behavior, enabling the system to test the validity of the analogy and sanction refinements where the analogy is incorrect. The system generates an envisionment of the scenario, which it then compares to the original observation. If the envisionment is consistent and complete with respect to the observation, then the explanation is considered successful. If it is inconsistent or fails to provide complete coverage, then revision is aimed at the points of discrepancy.

Revision. If an initial hypothesis fails, or an old hypothesis is inadequate for a new situation, an attempt is made to adapt it around points of inaccuracy. Revision relies on past experiences to guide the formation and selection of revision hypotheses. It considers behavior analogous to the current anomaly and considers how the current anomalous situation differs from prior situations that were consistently explained. This is the only component of PHINEAS that is not fully implemented.

3.1 Preference Criteria

An explanation system should focus on the most promising explanations first and provide a preferential ordering on fully developed hypotheses. A complete account of theory selection requires consideration of many complex factors, such as a theory's plausibility, coherence, effect on prior beliefs, simplicity, and specificity in accounting for the phenomenon. However, a number of important, more specific preference criteria are readily available and have been found useful in PHINEAS for establishing preference between competing hypotheses. These are:

C_{CE} *Conjectured entities.* Does the hypothesis conjecture the existence of a novel kind of entity, and if so, how many?

C_{VE} *Vocabulary extensions.* Does the hypothesis require the creation of new predicates, and if so, how many?

C_{CA} *Composite assumptions.* Does the hypothesis conjecture the existence of new physical processes or new knowledge structures (e.g., schemas, etc.), and if so, how many?

C_{AE} *Assumed entities.* Does the hypothesis assume the presence of a known type of entity not mentioned in the original scenario description, and if so, how many?

C_{AA} *Atomic assumptions.* Does the hypothesis make additional assumptions about the properties and interrelationships of objects in the scenario, and if so, how many?

The single preference criterion used to evaluate a hypothesis or compare two competing hypotheses is a function of these five metrics. They are ordered according to an approximate measure of decreasing "cost" and applied sequentially to prune the space of hypotheses:

$$\text{LEF} = (C_{CE},\ C_{VE},\ C_{CA},\ C_{AE},\ C_{AA})$$

Thus, an explanation that postulates the existence of a novel kind of entity (C_{CE}) is at all times deemed inferior to one that does not. Each criterion returns a number ($N \geq 0$) as described above, where a value of zero indicates success and a value greater than zero indicates failure. The function is used to select the most preferable explanation(s) from a given set as follows: First, each proposed explanation is evaluated by criterion C_{CE} and those that pass C_{CE} are retained. The process is repeated with the next criterion on the set of retained hypotheses until only a single hypothesis remains or the list of criteria is exhausted. If at any point all hypotheses evaluated by a particular criterion fail, the process stops and the current set is returned in increasing order according to their score, N, for that criterion.

This evaluative function produces an interesting property when viewed from the perspective of the four explanation scenarios described in Section 2:

Deductive scenario: This corresponds to explanations passing every criterion. It occurs when all of the antecedent features of the base are present in the target.

Assumption scenario: This corresponds to explanations passing every criterion but one of the last two, C_{AE} and C_{AA}. It occurs when some of the antecedent features of the base have no correspondent in the target, but may be consistently assumed to hold in the target.

Generalization scenario: This corresponds to explanations passing the first two criteria, C_{CE} and C_{VE}, but failing C_{CA}, in which a knowledge structure is viewed as "new" if it represents a modification of an existing knowledge structure. It occurs when some of the antecedent or consequent features of the base match an analogous set of features in the target, thus mapping the base theory to a situation beyond its declared scope.

Analogy scenario: This corresponds to explanations failing one of the first three criteria, C_{CE}, C_{VE}, or C_{CA}. It occurs when some of the features of the base match an analogous set of features in the target, or new vocabulary must be created to complete the mapping.

All four scenarios arise as a result of the same basic mechanism. The evaluative function causes PHINEAS to propose standard, deductive explanations if any are found. In their absence, conventional abductive explanations will be preferred. If existing theories are insufficient to provide an explanation, explanations adapting knowledge of potentially analogous phenomena will be offered. By using similarity as the single source for explanation generation, PHINEAS is able to offer a "best guess" in the presence of an imperfect or incomplete domain theory.

4 Examples of PHINEAS' Behavior

PHINEAS has been tested on over a dozen examples including explanations of evaporation by analogy to boiling, liquid flow, and dissolving; torsional and LC circuit oscillators by analogy to a spring-mass harmonic oscillator; osmosis by analogy to liquid flow; and floating of a balloon by analogy to an object floating in water. For example, when only given knowledge of liquid flow, the system is able to interpret the three situations shown in Figure 1:

(a) A beaker contains more water than a vial to which it is connected by an unknown object. Why does the water level in the beaker decrease and the water level in the vial increase?

(b) Two containers sharing a common wall of unknown substance each hold some solution. Why does one solution's level decrease and concentration increase while the other solution's level increases and concentration decreases?

(c) What causes a hot brick and cold water to change to the same median temperature when the brick is immersed in the water?

In each case, PHINEAS bases its explanation on the case's similarity to liquid flow. In the first, the phenomenon most similar to an observation of liquid flow is liquid flow itself, thus suggesting that the unknown object may be a fluid path. In this work, identicality is viewed as an extreme form of similarity. The second behavior, called *osmosis*, represents a close generalization of liquid flow when viewed as flow of solute under osmotic pressure through a selective kind of fluid path. In the final "heat flow" observation, PHINEAS draws an across-domain analogy to liquid flow phenomena and conjectures the existence of a new type of fluid that affects an object's temperature. Its predictions based on this new heat flow model are shown in Figure 2. All three interpretations are produced by a single mechanism that forms its explanations from theories about phenomena most similar to the current situation.

5 Discussion

This paper has described a unified, similarity-driven method for explanation that seeks the best match between an observation to be explained and understood phenomena. This addresses imperfect theory problems by enabling matching of analogous rather than identical features, reducing the need to have a precisely defined set

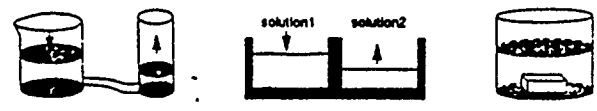
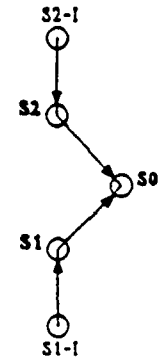


Figure 1: Three phenomena that PHINEAS explains by their similarity to liquid flow: (a) liquid flow, (b) osmosis, (c) heat flow.

Quantity	S2-I	S2	S0	S1-I	S1
D$_s$[FLOW-RATE(PI0)]	-	-	-	-1	-1
D$_s$[FLOW-RATE(PI1)]	-1	-1	-	-	-
D$_s$[AMOUNT-OF(SK-CS-WATER-BEAKER-1)]	-1	-1	0	1	1
D$_s$[AMOUNT-OF(SK-CS-WATER-VIAL-1)]	1	1	0	-1	-1
D$_s$[PRESSURE(SK-CS-WATER-BEAKER-1)]	-1	-1	0	1	1
D$_s$[PRESSURE(SK-CS-WATER-VIAL-1)]	1	1	0	-1	-1
D$_s$[TEMPERATURE-IN(BRICK)]	-1	-1	0	1	1
D$_s$[TEMPERATURE-IN(WATER1)]	1	1	0	-1	-1
A[AMOUNT-OF(SK-CS-WATER-BEAKER-1)]	>0	>0	>0	>0	=0
A[AMOUNT-OF(SK-CS-WATER-VIAL-1)]	=0	>0	>0	>0	>0
A[TEMPERATURE-IN(BRICK)] A[TEMPERATURE-IN(WATER1)]	>	>	=	<	<
ACTIVE(PI0)	F	F	F	T	T
ACTIVE(PI1)	T	T	F	F	F



Processes:
PI0 PROCESS-1(SK-WATER-1 WATER1 SK-CS-WATER-VIAL-1 BRICK SK-CS-WATER-BEAKER-1 (COMMON-FACE BRICK WATER1))
PI1 PROCESS-1(SK-WATER-1 BRICK SK-CS-WATER-BEAKER-1 WATER1 SK-CS-WATER-VIAL-1 (COMMON-FACE BRICK WATER1))

Figure 2: Envisionment produced by the hypothesized caloric model when applied to the brick immersed in water scenario. States are distinguished only by derivative and process values. They are split when this distinction produces a state lasting an interval of time (S2) and also lasting for an instant (S2-I).

of necessary and sufficient conditions for each theory, and enabling knowledge of a familiar domain to aid reasoning about another domain. Importantly, all explanations are formed with a single mechanism, with distinctions between deductive, abductive, and novel analogical explanations arising out of the evaluation process. Initial viability of the method has been demonstrated by **PHINEAS** on a variety of complex examples.

Explanation systems rarely address problems due to lack of applicable knowledge about the domain. However, there are a few exceptions in addition to **PHINEAS**. These include using knowledge of abstract patterns of causality (Pazzani, 1987), experimentation (Rajamoney, 1990), and meta-theoretic rules enabling abduction to include assumption of new causal rules (O'Rorke et al, 1990). This work shares much of the philosophy behind case-based reasoning, which uses similar past problem solving experiences to solve new cases (Kolodner el al, 1985; Hammond, 1989). Motivated by the complex causal reasoning task and the concern with across-domain analogies, this work provides a more sophisticated notion of analogical similarity and require a deep causal analysis of the consistency of a hypothesis.

Acknowledgements

This paper is an excerpt from (Falkenhainer, 1990). This work has benefited from discussions with Ken Forbus, Dedre Gentner, John Collins, and Dennis DeCoste. Jeff Shrager and Pat Langley provided valuable comments on prior drafts.

References

Falkenhainer, B. (1988). *Learning from Physical Analogies: A Study in Analogy and the Explanation Process.* PhD thesis, University of Illinois at Urbana-Champaign.

Falkenhainer, B. (1990). A unified approach to explanation and theory formation. In J Shrager and P Langley (Eds.), *Computational models of discovery and theory formation. (in press).*

Falkenhainer, B, Forbus, K. D, and Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence, 41,* 1-63. (University of Illinois Technical Report UIUCDCS-R-87-1361, July 1987).

Forbus, K. D. (1984). Qualitative process theory. *Artificial Intelligence, 24.*

Gentner, D. (1983). Structure-Mapping: A theoretical framework for analogy. *Cognitive Science* 7(2), 155-170.

Hammond, K. J. (1989). *Case-based planning: Viewing planning as a memory task.* Boston, MA: Academic Press.

Kolodner, J. (1984). *Retrieval and organizational strategies in conceptual memory: A computer model.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Kolodner, J, Simpson, R. L, and Sycara-Cyranski, K. (1985). A process model of cased-based reasoning in problem solving. IJCAI-85.

O'Rorke, P, Morris, S, and Schulenburg, D. (1990). Theory formation by abduction: Initial results of a case study based on the chemical revolution. In J Shrager and P Langley (Eds.), *Computational Models of Discovery and Theory Formation. (in press).*

Pazzani, M. J. (1987). Inducing causal and social theories: A prerequisite for explanation-based learning. *MLW-87.*

Rajamoney, S. (1990). Towards a computational model of theory revision. In J Shrager and P Langley (Eds.), *Computational Models of Discovery and Theory Formation. (in press).*

On the "Logical Form" of Abduction

John R. Josephson

Laboratory for Artificial Intelligence Research
Department of Computer and Information Science and
Center for Cognitive Science
228 Bolz Hall, 2036 Neil Avenue
The Ohio State University
Columbus, Ohio 43210

Introduction: Evidence, Inference, and Justification

I take it that there is a distinctive kind of evidential support which follows a pattern pretty nearly as follows:

> D is a collection of data (facts, observations, givens).
> H explains D (would, if true, explain D).
> No other hypothesis is able to explain D as well as H does.
>
> ---
>
> Therefore, H is probably true.

This is the pattern I identify with the term "abduction." Really, when we want to be precise, we should distinguish "abductive support" (for an evidence relationship), "inference to the best explanation", "abductive inference," and "best-explanation reasoning" (for inference processes), and "abductive justification" for an appeal to evidence to support a conclusion. Three faces of abduction: evidence, inference, and justification.

The judgment of likelihood associated with an abductive conclusion should depend on the following considerations (and it typically does in the inferences we actually make):

- how decisively H surpasses the alternatives,
- how good H is by itself, independently of considering the alternatives (e.g. we should be cautious about accepting a hypothesis, even if it is clearly the best one we have, if it is not sufficiently plausible in itself),
- judgments of the reliability of the data, and
- how much confidence there is that all plausible explanations have been considered (how thorough was the search for alternative explanations).

Beyond the judgment of likelihood, willingness to accept the conclusion should (and typically does) depend on:

- pragmatic considerations, including the costs of being wrong and the benefits of being right,
- how strong the need is to come to a conclusion at all, especially considering the possibility of seeking further evidence before deciding.

The core intuition is that a body of data provides evidence for a hypothesis that satisfactorily explains or accounts for the data (or at least it provides evidence if that hypothesis is better than other hypotheses).

Of course it remains to elucidate what makes a hypothesis explanatory relative to some presumed fact, and what makes one explanation better than another.

The relationship between a body of given facts, and some conclusion for which those facts gives evidence, may be deductive whether we know it or not. More precisely, the statements of some set of facts might deductively entail certain other statements of fact, whether or not we are aware of that entailment. We might not happen to make the inference from one to the other, even though we would be logically justified if we were to do so. Thus evidence relationships can be considered to exist as a matter of objective (logical) fact, apart from the making of actual inferences, and apart from actual attempts to justify one's conclusions.

In this short paper I assume without argument that abductive support really exists, that inference to the best explanation is common in ordinary life (and in diagnostic reasoning, etc.), and that abductive justifications are commonly appealed to in ordinary life, in the law courts, and in the justifications for scientific conclusions.

Deduction and Abduction

Deductions support their conclusions in such a way that the conclusions must be true, given the premises; they convey conclusive evidence. Other forms of evidential support are not so strong, and though significant support for a conclusion may be given, possibility of error remains. Abductions are typically of this kind.

To a great degree the patterns of (valid) deductive inference have been well characterized by formal logic, from the syllogistic logic of Aristotle, through modern mathematical logic. Yet despite the great successes of modern formal logic, especially in capturing the forms of justification that occur in mathematical proofs (but see Goodman (1987)), it is nevertheless not correct to think that all forms of deductive inference have been satisfactorily analysed. Deductive logic is not a finished science. A worse mistake is to completely identify valid deductive inference with one particular mathematical system, such as First Order Predicate Calculus.

Consider the following logical form, commonly called "disjunctive syllogism."

$P \vee Q \vee R \vee S \vee \cdots.$
But $\neg Q, \neg R, \neg S, \neg \cdots.$

Therefore, P.

This form is deductively valid. Moreover, the abduction schema fits this form, if we assert that we have exhaustively enumerated all of the possible explanations for the data, and that all but one of the alternative explanations has been decisively ruled out. Typically, however, we will have reasons to believe that we have considered all plausible explanations (i.e. those which have a significant chance of being true), but these reasons stop short of being conclusive. For example we may have struggled to formulate a wide variety of possible explanations, but cannot be sure we have covered all plausibles. Under these circumstances the we can assert a proposition of the form of the first premise, but assert it only with a kind of qualified confidence. Typically, too, alternative explanations can be discounted for one reason or another, but not decisively ruled out. The lesson to be drawn is that abductive inferences can actually under some circumstances be valid deductive inferences, and that abductions are deductive in the limit.

It is also worth noting that abductions have an interesting way of turning negative evidence (against some hypotheses) into positive evidence (for alternative explanations).

Another apparent connection between deduction and abduction occurs if one tries to give a deductive account of explanation. There have been two main traditional attempts to analyse explanations as deductive proofs, neither attempt particularly successful (at least in my judgment). Aristotle maintained that an explanation is a syllogism of a certain form (Aristotle 1941) (actually c. 330 bc) that also satisfies various (informal) conditions, one of which is that the middle term is the cause of the thing to be explained. More recently (considerably) Hempel has modernized the logic and proposed the "covering law" or "deductive nomological" model of explanation (Hempel 1965) . For a brief summary of deductive and other models of explanation see (Bhaskar 1981) . The main difficulties with these accounts (besides Hempel confounding the question of what makes an ideally good explanation with the question of what it is to explain at all) is that being a proof appears to be neither necessary nor sufficient for being an explanation. Consider the following:

> Why does he have burns on his hand? Explanation: He sneezed while cooking pasta and upset the pot.

The point of this example is that an explanation is given, but no proof; and while it could be *turned into* a proof by including additional propositions, this would amount to gratuitously completing what is on the face of it an incomplete explanation. Under the circumstances (incompletely specified), sneezing and upsetting the pot were presumably *causally sufficient* for the effect, but that is quite different from being *logically sufficient*. The case that explanations are not necessarily proofs becomes even stronger if we consider psychological explanations, and explanations which are fundamentally statistical (e.g. where quantum phenomena are involved), since it is clear that causal determinism cannot be assumed, and so the antecedent conditions cannot be assumed to be even causally sufficient for the effects.

Conversely, there are many proofs which fail to be explanations of anything, for example in classical mechanics an earlier state of a system can be deduced from a subsequent state, but the earlier state cannot be said to be explained thereby. Also note that P can be deduced from $P \wedge Q$, but is not explained thereby.

Thus I conclude that explanations are not proofs in any particularly interesting sense (of course they can always be PUT as proofs; the point is that this does not succeed in capturing anything essential or especially

useful.)

Abductions are (or can be) truth producing, that is, at the end of an abductive process, having accepted a best explanation, we may have more information than we knew before. The abduction, so to speak, stands upon the old information of its premises, and makes new information not previously encoded there at all. This can be contrasted with deductions, which can be thought of as extracting out explicitly in their conclusions, information that was already implicitly contained in the premises. While abductions do not typically offer the truth-preserving certainty of valid deductions, and are almost always accompanied by some degree of doubt, they are capable of accomplishing something else that deductions cannot, namely the introduction of new vocabulary. Valid deductive inferences cannot contain terms in their conclusion that do not occur in their premises. Abductions can "interpret" the given data in a new vocabulary.

Abductions can display "emergent certainty," that is, the conclusion of an abduction can have, and be deserving of, more certainty than *any* of its premises. This is unlike a deduction, which is no stronger than the weakest of its links (though separate deductions can converge for parallel support). For example I may be more sure of the bear's hostile intent, than of any of the details of its hostile gestures; I may be more sure of the meaning of the sentence, than of my initial identifications of any of the words; more sure of the overall theory, than of the reliability of any of the single experiments on which it is based.

In summary we may say that deductions are "truth-conserving", while abductions are "truth-producing".

Causality and Abduction

The relationship of explainer to explained is better described as "causes" than as "implies." An explanation is an assignment of causal responsibility; it tells a causal story. (At least this is the sense of the term "explanation" relevant for abduction.) Thus abduction is basically a process of reasoning from effect to cause. Finding possible explanations is finding certain interesting possible causes of the thing to be explained. (There are apparent counterexamples to this view, but I claim they all trade on an overly narrow view of causation. "Cause" in this context must be understood somewhat more broadly than its usual modern senses of "mechanical" or "efficient" or "event-event" causation. There is not room to argue that here, however.) For a well-developed historical account of the connections between causality and explanation see (Wallace 1972, 1974).

Inductive Generalization and Abduction

Harman (1965) argued that "inference to the best explanation" (IBE) is *the* basic form of non-deductive inference, subsuming "enumerative induction" and all other forms of non-deductive inference. He argued quite convincingly that IBE is a common and important pattern of inference, and that it subsumes sample-to-population inferences, i.e., inductive generalizations, as a special case. [This is my way of putting the matter.] The weakness of his overall argument was that other forms of non-deductive inference are not seemingly subsumed by IBE, most notably population-to-sample inferences, i.e., predictions. The main problem is that the conclusion of a prediction does not seem to explain anything. Nevertheless Harman's basic argument (suitably augmented) seems quite sound, if the conclusion is restricted to be that inductive generalizations are a special class of abductions. (See Josephson (1982) pp. 107-130 for more details.)

Probabilities and Abduction

Bayes's Theorem can be viewed as a way of describing how simple alternative causal hypotheses can be weighed. Thus, if suitable knowledge of probabilities is available, the mathematical theory of probabilities can, in principle, guide our abductive evaluation of explanatory hypotheses to determine which is best. In practice, however, it seems that rough qualitative confidence levels on the hypotheses are enough to support abductions, which then produce rough qualitative confidence levels for their conclusions. It is certainly possible to model these confidences as continuous, and on rare occasions one can actually get knowledge of numerical confidences (e.g. for Blackjack), but for the most part numerical confidences are unavailable and unnecessary for reasoning. People are good abductive reasoners in the absence of close estimates of confidence. In fact it seems that, if confidences need to be estimated closely, then it must be that the best hypothesis is not much better than the next best, in which case no conclusion can be confidently drawn. (Recall the condition, mentioned earlier, that the confidence of an abductive conclusion depends on how decisively the best explanation surpasses the alternatives.) Thus it appears that confident abductions are possible only if confidences do not need to be estimated closely!

Furthermore it appears that accurate knowledge of probabilities is *not* commonly available, because the

probability associated with an expected possible event is not even well-defined. This is what I have been calling "the scandal of probabilities." There is almost always a certain arbitrariness about which reference class is chosen to base the probabilities, the larger the reference class the more reliable the statistics, but then the less relevant they are; while the more specific the class, the more relevant, but the less reliable. Is the likelihood that the next patient has the flu best estimated based on the frequency in all of the people in the world over the entire history of medicine? It seems better to at least control for the season and narrow the class to include just people at this particular time of the year. (But then causal understanding is starting to creep into the considerations, but that isn't probabilities.) Furthermore each flu season is somewhat different, so we would be best to narrow to just considering people THIS year. Then of course the average *patient* is not the same as the average *person*, etc., etc., so the class should probably be narrowed further to something like: people who have come LATELY to doctors of this sort, of this particular age, race, gender, and social status. Now the only way the doctor could have statistics that specific, would be to rely on his or her own most recent experience, which would only allow for very rough estimates of likelihood. There is a Heisenberg-like uncertainty about the whole thing - the closer you try to measure them, the rougher the numbers get. The conclusion to draw is that using real numbers for confidence levels is misplaced precision. "In general the problem faced by intelligence isn't reasoning with uncertainty, but reasoning *despite* uncertainty." (Chandrasekaran 1987). That is, Even If we could define some ideal reasoner who worked completely rationally on the basis of probabilities, and Even If strategies could be devised that would make it possible to actually feasibly make all the computations, then it Would Be Wasted Effort anyway, because almost all of the numbers would only be rough approximations, and all that would still have to be translated into Tentative Categorical Judgments in order to support hypothetical reasoning and action.

Abduction as an Inference Process

An abductive process aims at a satisfactory explanation, one that can be confidently believed (accepted into memory). It might, however, be accompanied in the end with some explicit qualifications, for example some degree of assurance, or some doubt. (One main form of doubt is just hesitation from being aware of the possibility of alternative explanations.) Along the way an abductive process might seek further information beyond that which is presupposed in the data initially to be explained. For example there may be a need to distinguish between explanatory alternatives, or for help in forming hypotheses, or help in evaluating them. Thus often abductive processes are not immediately concluded, but rather suspend to wait for answers to information-seeking questions.

Humans can understand sentences, form little causal theories of everyday events, and so on, apparently performing complex abductive inferences very quickly, even in fractions of a second. Yet when we set out to form a hypothesis for some body of data, we have in general no advance assurance that the best explanation will turn out to be a simple hypothesis. In fact it is typical that an abductive conclusion be a multi-part composite hypothesis, with the parts playing differing roles in explaining different parts of the data. For example the meaning of a sentence must be some kind of composite hypothesis, formed on the fly as the sentence is understood, including components that function to explain the word order, choice of vocabulary, intonation, and so on.

It is not in general a computationally feasible strategy for finding the best explanation for a given body of data to consider all possible combinations of elementary hypotheses, comparing each composite hypothesis with each to see which is the best. It would be better not to need to explicitly generate all of the combinations, since the number of them is an exponential function of the number of elementary hypotheses available, and it rapidly becomes an impractical computation unless almost all elementary hypotheses can be ruled out in advance. Thus a general strategy for abduction must avoid generating more than a small number of composite hypotheses, either by ruling out all but a few elementary hypotheses, or by generating a small number of composites by methods that implicitly compare those generated to the large number of those that are not.

We may hypothesize that the functional needs of abductive information processing are similar across widely different domains. If this is so, then there may be a single generic architecture for the generic information processing task of forming a confident explanation (if possible) for a given body of data. (Or perhaps there are a small number of such architectures.) In fact I have proposed elsewhere (Josephson 1989) that at a certain level of description both "deliberative," and "compiled" or "perceptual" abductions, can be accommodated by a single architecture, and thus that the information processing that occurs in diagnosis, story understanding, vision, scientific theory formation, hearing, understanding spoken language, and so on, are all accomplished

by variations, incomplete realizations, or compilations (domain-specific optimizations) of one basic computational mechanism.

However this claim to extreme generality turns out, my colleagues and I at Ohio State have been developing a series of generic mechanisms for abductive processing, and at least some degree of generality has already been achieved (Josephson et al. 1987) (Goel, Sadayapan, and Josephson 1988) (Punch at al. 1989) .

Acknowledgments

Thanks to B. Chandrasekaran for providing the intellectual environment within which these investigations into abductive processing have taken place, and for many fruitful discussions. Thanks besides to Michael C. Tanner, Ashok Goel, Dean Allemang, William Punch, Tom Bylander, Jack Smith, Jr., Susan Josephson, Jon Sticklen, Michael Weintraub, Todd Johnson, Richard Fox, Hari Narayanan, Diana Smetters, Terry Patten, and Jordan Pollack for many stimulating discussions of abductive processing and of these issues. My apologies to whomever I am forgetting. The errors that remain are of course mine, especially the exorbitant claims. My research on abduction has been supported by the National Heart, Lung and Blood Institute, NIH Grant 1 R01 HL 38776-01, by the Defense Advanced Research Projects Agency under RADC contract F30602-85-C-0010, by DARPA and The Air Force Office of Scientific Research under contract F49620-89-C-0110, and by DARPA and the National Science Foundation under grant CBT-8703745.

References

Aristotle (1941). Posterior analytics. In McKeon, R., editor, *The Basic Works of Aristotle*, pages 110–186. Random House, New York, N.Y. Translated by G.R.G. Mure.

Bhaskar, R. (1981). Explanation. In Bynum, W., Browne, E., and Porter, R., editors, *Dictionary of the History of Science*, pages 140–142. Princeton University Press, Princeton, NJ.

Chandrasekaran, B. (1986). Generic tasks in knowledge-based reasoning: High-level building blocks for expert system design. *IEEE Expert*, pages 23–30. Fall 1986.

Chandrasekaran, B. (1987). on reasoning despite uncertainty. verbal communication.

Goel, A., Sadayappan, P., and Josephson, J. R. (1988). Concurrent synthesis of composite explanatory hypotheses. In *Proceedings of the Seventeenth International Conference on Parallel Processing*, pages 156–160.

Goodman, N. (1987). Intensions, church's thesis, and the formalization of mathematics. *The Notre Dame Journal of Formal Logic*, 28:473–489.

Harman, G. (1965). The inference to the best explanation. *Philosophical Review*, LXXIV:88–95.

Hempel, C. G. (1965). *Aspects of Scientific Explanation*. The Free Press, New York.

Josephson, J. R. (1982). *Explanation and Induction*. PhD thesis, The Ohio State University.

Josephson, J. R. (1988). Reducing uncertainty by using explanatory relationships. In *Proceedings of the Space Operations Automation and Robotics Conference-1988 (Soar-88)*, pages 149–151.

Josephson, J. R. (1989). A layered abduction model of perception: integrating bottom-up and top-down processing in a multi-sense agent. In *Proceedings of the NASA Conference on Space Telerobotics*. to appear.

Josephson, J. R., Chandrasekaran, B., Smith, Jr., J., and Tanner, M. C. (1987). A mechanism for forming composite explanatory hypotheses. *IEEE Transactions on Systems, Man and Cybernetics, Special Issue on Causal and Strategic Aspects of Diagnostic Reasoning*, SMC-17(3):445–54.

Punch, W., Tanner, M., Josephson, J., and Smith, J. (1989). Using the tool peirce to represent the goal structure of abductive reasoning. Technical report, Michigan State University and The Ohio State University.

Tanner, M. and Josephson, J. (1988). Abductive justification. Technical report, Ohio State University Laboratory for AI Research.

Wallace, W. A. (1972). *Causality and Scientific Explanation*, volume 1, Medieval and Early Classical Science. University of Michigan Press, Ann Arbor, MI.

Wallace, W. A. (1974). *Causality and Scientific Explanation*, volume 2, Classical and Contemporary Science. University of Michigan Press, Ann Arbor, MI.

A quick review of hypothetical reasoning based on abduction

Randy Goebel

Department of Computing Science
University of Alberta
Edmonton, Alberta, Canada T6G 2H1
goebel@cs.ualberta.ca

Abstract

Abduction is an unsound rule of inference of the form "from P and $Q \rightarrow P$, infer Q." To emphasize the distinction between abduction and *sound* rules of inference, Q is called a *hypothesis*, thus abduction provides the basis for hypothetical reasoning systems.

We describe a simple system of hypothetical reasoning whose essentials are familiar to anyone who has analyzed the reasoning strategy of INTERNIST, worked on residue resolution, done any kind of "explanation-based" x where x is "learning," "reasoning," "concept formation," etc. We then provide terse descriptions of the system's relationship to deduction, induction, analogy, probabilistic reasoning, and nonmonotonic reasoning.

1 Introduction

Abduction is an unsound rule of inference of the form "from P and $Q \rightarrow P$, infer Q." To emphasize the distinction between abduction and *sound* rules of inference, Q is called a *hypothesis*, thus abduction provides the basis for hypothetical reasoning systems.

Whether a reasoning system is actually doing abduction or not has a lot to do with the form of the theory from which reasoning proceeds. For example, the abductive rule of inference is indistinguishable from any attempt to constuct an ordinary first order deductive proof in a goal-directed way. From the theory Q, $Q \leftarrow P$, P, a goal-directed deductive proof of Q necessarily proceeds to the subgoal P by using what amounts to the abductive rule of inference; a completed derivation relies on the subgoal (hypothesis) being directly derivable from the initial theory.

So, if goal-directed theorm-provers use abduction to create subgoals, what distinguishes deductive theorem-proving from abductive or hypothetical theorem-proving? Of course it is the truth-conditional status of the subgoals or hypotheses. In deduction, subgoals in a successful derivation have their truth grounded in the original theory; in hypothetical reasoning, the hypotheses are not deductively established.

The following section provides a thumbnail sketch of a simple hypothetical reasoning system based on first order clausal logic without equality. It is "hypothetical" or "abductive" because it explicitly distinguishes formulas whose truth conditions are established (either by asserting them as axioms or demonstrating their deductive derivation) from those which are not. The latter are called hypotheses, which can participate in any derivation as long as there is *sufficient evidence* to assume them. As we shall discuss, the notion of what constitutes sufficient evidence is a central issue in drawing the boundaries between different kinds of hypothetical reasoning.

2 A brief description of a hypothetical reasoning system

Using Prolog as a systems programming language, we have implemented many variations of the following hypothetical reasoning system specification[GFP86, GG87, Poo88].

The logic is full first order clausal logic without equality; the proof theory is a goal-directed theorem prover based on Prolog but modified with Loveland's MESON proof procedure idea to get a full clausal prover (see [PGA87] for details).

The essence of the abuctive component arises from a simple distinction in any applied theory every database DB of sentences is separated into facts F and hypotheses H. In the goal-directed search for a derivation of an alleged theorem (goal) G, an ordinary deductive proof is attempted using sentences from both F and H. If the derivation succeeds without using any sentences from H, then the derivation is wholly deductive and G is a logical consequence of F

Otherwise, the participation of any sentences from H must be accompanied by some kind of rational support (cf [Isr80, Isr83, Isr88]). Although the extra evidence in support of hypotheses varies from implementation to implementation (see below), all systems insist that any member of H must be consistent with the existing facts F and all other hypotheses participating in a derivation.[1]

Before continuing to address the relationship to other styles of reasoning, we conclude with a simple example. Let

$$F = \{bird(tweety), emu(X) \rightarrow \neg flies\},$$
$$H = \{bird(X) \rightarrow flies(X)\},$$
$$G = flies(tweety).$$

Note that all variables are universally quantified. The appropriate derivation is

$$flies(tweety)$$
$$flies(tweety) \leftarrow bird(tweety)$$
$$bird(tweety)$$

and the use of the instance of the hypothesis is supported by verifying that F does not derive the negation of the hypothesis, i.e., it is consistent. We name the instance of the hypothesis as the "theory" T that explains the goal G. This use of the word "theory" is consistent with the notion of nomological explanation (e.g., [Hem65]).

Note that this style of reasoning is nonmonotonic, as augmenting F with $emu(tweety)$ will have $\neg flies(tweety)$ as a consequence, but will not support $flies(tweety)$ as the required $T = \{bird(tweety) \rightarrow flies(tweety)\}$ is not consistent with the augmented F.

3 Relationship to deduction

As suggested above, it is easy to confuse some forms of deduction with abduction because the backward or goal-directed application of modus ponens is indistinguishable from the abductive rule of inference. The simplest way to state the relationship between the two is that the truth-conditional status of abduced sentences is not established in any deductive way, and so must be supported by a variety of rationalizations, including consistency, lack of evidence to the contrary, high probability, etc.

[1] Poole has shown how this specification provides sufficient expressive power for default reasoning [Poo88]; Lin and Goebel show how Przymusinski's circumscriptive theorem proving [Prz89] is equivalent to a conservative form of prediction, based on this hypothetical reasoning system.

4 Relationship to induction

The simplest description of induction is typically given as the rule of universal generalization, as follows: from $p(a)$, infer $p(X)$, for all X. To see the relationship to abduction, consider the situation with $G = p(a)$, $F = \{\}$ and H consists of all sentences in the language (or at least a generator that produces the potentially infinite set in some well-defined order). The hypothetical reasoning system is doing induction whenever it constructs derivations of $p(a)$ that use instances of any sentence of H that is more general than $p(a)$. For example, selecting $p(X)$ from H to create the explanation $T = \{p(a)\}$ is boring, but it is still induction. We might consider the restriction to more "interesting" hypotheses, e.g., $p(X) \leftarrow q(X), q(X)$ if we have any rational well-defined reason for doing so. It is already well-known that fabricating rational inductive schemas (i.e., waiting for the appropriate member of H to be generated) is not a simple problem.

5 Relationship to analogy

If you believe that analogical reasoning is not deductive (some don't, e.g., [DR87]), then perhaps it is abductive? Consider

$$F = \{p(a)\}$$
$$H = \{X = Y\}$$
$$G = p(b)$$

Since everyone *will* admit that analogy reasoning is somehow related to some kind of equality (e.g., partial equality, partial relevant equality, etc.), we might simply treat various kinds of equality definitions as hypotheses. In this case, $p(b)$ can be derived if we assume $T = \{a = b\}$ as an instance of the hypothesis $X = Y$, i.e., $F \cup \{a = b\} \models p(b)$. As shown in [Goe89], there are plenty of possible equality definitions possible. The essential relationship suggests that we are doing analogical reasoning whenever we do abductive reasoning that involves assuming something about some kind of equality.

In the above example, so called "source" and "target" knowledge consists only of the single fact $p(a)$, The only similarity assumption is the hypothesis schema $H = \{X = Y\}$, which we interpret to mean that any X is equal to any Y. In the vocabulary of analogical reasoning, we have concluded $p(b)$ on the basis of source knowledge $p(a)$ and mapping $a = b$. The theory T is admittedly rather weak evidence for the "analogical conclusion," but the general structure of hypothetical reasoning provides for arbitrarily complex justifications of arbitrary theories (e.g.,

see [GG89]). The simple point is that assumptions about equality are the substance from which mappings are made; theory preference is a generalization of the methods developed for ranking analogical mappings (e.g., [Hal89]).

Note further, that this example has no explicit axioms for equality (other than the hypothesis). Because of this, both the goal and the fact must be written according to the standard transformation that explicitly introduces equality (cf. [Cla78, LvE84]), viz.

$p(a)$ becomes $p(X) \wedge X = a$
the goal $p(b)$ becomes $p(X) \wedge X = b$

where variables in goals are existentially quantified. In this way the derivation can be constructed without the explicit need for a full equality axiomatization. This is not fully general, as shown in [Goe89].

Also note that the predicate symbol = is often assumed to mean identity in an interpretation, as opposed to an arbitrary equivalence relation. Many theories of nonmonotonic reasoning that exploit model-theoretic minimization (cf. [Rei87]) assume unique names axioms, viz.

$$\alpha_1 \neq \alpha_2 \wedge \alpha_2 \neq \alpha_3 \ldots$$

where each α_i is an individual constant of the language, and the intended interpretation of the symbol = is identity. Usually, because of the proof-theoretic difficulties presented by equality axioms, this use of identity is not explicit. This specification of analogical reasoning based on abduction suggests an intuitive interpretation of the symbol = is something like "sufficiently similar, according to what is currently known."

6 Relationship to probability

Since hypothetical reasoning is nonmonotonic, it is not surprising that there is a relationship between abduction and probability. While the debate over probability versus logic has been somewhat polarized (e.g., [Me88]), it is simple to argue that one way to address the rational choice of competing explanations is to choose the most probable. For example, if we have

$F = \{p(a), p(b)\}$
$H = \{X = X\}$
$G = p(c)$

we get two explanations $T_1 = \{a = c\}$, and $T_2 = \{b = c\}$. It is desirable to have a probability measure such that $P(a = c|F)$ was different from $P(b = c|F)$ in order to rank explanations.

So abduction can produce explanations and rely on some system of conditional probability to help choose the best explanation. Two different approaches to this amalgamation are provided in [LG89] and [GG89]. The former uses a monadic first order language in graphical form, and attributes a probability value to each link in the network (cf. [Pea86]). Reasonable assumptions about independence and localization of probabilistic influences allow abduction to the most probable explanation in time that is exponential in the number of atomic goals (observations), which are usually few. The algorithm uses a Steiner tree algorithm to find explanations, and incrementally computes the accumulating probabilities during construction of the explanation.

The latter approach, [GG89] uses hypothetical reasoning directly, together with an extended metalanguage for expressing independence relations on categories of events (predicates). Inference with the independence statements is combined with explanation generation to identify those most likely.

7 Relationship to nonmonotonic reasoning

As illustrated above, hypothetical reasoning based on abduction is nonmonotonic. The nonmonotonicity arise because the set of explanations from a fixed H is not montonic with respect to a montonically enlarging F. As illustrated in the Tweety example in Section 2, new facts invalidate previously determined explanations.

It is likely that there is no domain independent strategy for selecting the best T (cf. [Ale88, DW89]), so the best one can hope for is to build systems that find the best domain dependent Ts as efficiently as possible. In this regard, we have investigated ways in which deductive attempts to establish consistency are related to truth maintenance systems(TMS), constraint programming and general techniques for improving the efficiency of theorem provers [SG89]. Note that, despite the pessimism as regards the undecidability of truth maintenance systems for first order systems (e.g., [RJ87]), we have constructed such systems that are empirically more efficient than DeKleer-style TMSs. Our empirical improvements, motivated by the impossibility of asymptotic analysis, exploit the properties of the finite failure derivation tree developed in attempts to establish hypothesis consistency.

8 Conclusion

Abduction is a logical method of isolating interesting hypotheses, and so is naturally applicable in every situation where goal-directed reasoning proceeds within the context of uncertain or incomplete information. Our research strategy has been to attack such problems with the undecidable version of a hypothetical reasoning specification based on abduction, and to empirically determine necessary improvements in both specification and implementation, for various applications.

Acknowledgements

These ideas derive from years of interaction with various people including those at the University of BC, the University of Waterloo, the University of Alberta. I doubt, however, that any one of them would completely agree with this rather broad statement of the nature of abductive reasoning. This work was has been supported by Natural Sciences and Engineering Research Council of Canada grants A0894, G1587, OGP9443, and INF36861.

References

[Ale88] R. Aleliunas. Comments on peter cheeseman's *an inquiry into computer understanding*. *Computational Intelligence*, 4(1):67–69, 1988.

[Cla78] K.L. Clark Negation as failure. In H. Gallaire and J. Minker, editors, *Logic and Data abases*, pages 293–322. Plenum Press, New York, 1978.

[DR87] T.R. Davies and S.J. Russell. A logical approach to reasoning by analogy. In *Proceedings of IJCAI-87*, pages 264–270, Milan, Italy, August 23-28 1987.

[DW89] J. Doyle and M. Wellman. Impediments to universal preference-based default theories. In *Proceedings of KR89*, pages 94–102, Toronto, Ontario, May 15-18 1989.

[GFP86] R. Goebel, K. Furukawa, and D. Poole. Using definite clauses and integrity constraints as the basis for a theory formation approach to diagnostic reasoning. In *Proceedings of the Third International Conference on Logic Programming*, pages 211–222, London, England, July 14-18 1986. Imperial College.

[GG87] R. Goebel and S.D. Goodwin. Applying theory formation to the planning problem In *Proceedings of the AAAI Workshop on The Frame Problem in Artifical Intelligence*, pages 207–232, Lawrence, Kansas, April 12-15 1987.

[GG89] S.D. Goodwin and R. Goebel. Statistically motivate defaults. (manuscript), 1989.

[Goe89] R. Goebel. A sketch of analogy as reasoning with equality hypotheses. In K Jantke, editor, *Analogical and Inductive Inference*, volume 397 of *Lecture Notes in Computer Science*, pages 243–253. Springer Verlag, Berlin, 1989.

[Hal89] R.P. Hall. Computational approaches to analogical reasoning: A comparative analysis. *Artifical Intelligence*, 39(1)·39–120, 1989.

[Hem65] C.G. Hempel. *Aspects of scientific explanation and other essays in the philosophy of science*. The Free Press, New York, 1965.

[Isr80] D.J. Israel. What's wrong with nonmonotonic logic? In *Proceedings of AAAI-80*, pages 99–101, Stanford, California, August 18-21 1980. Stanford University.

[Isr83] D.J. Israel. The role of logic in knowledge representation. *Computer*, 16(10):37–41, 1983.

[Isr88] D.J. Israel. On cheeseman. *Computational Intelligence*, 4(1):85–86, 1988.

[LG89] D. Lin and R. Goebel. A probabilistic theory of abductive diagnostic reasoning. Research report TR 89-25, Department of Computing Science, University of Alberta, Edmonton, Canada, October 1989.

[LvE84] J.W. Lloyd and M.H. van Emden. A logical reconstruction of Prolog ii. In *Proceedings of the Second International Logic Programming Conference*, pages 115–125, Uppsala, Sweden, July 2-6 1984. Uppsala University.

[Me88] M. McLeish (ed.). Taking issue: Cheeseman's *an inquiry into computer understanding*. *Computational Intelligence*, 4(1):57–142, 1988.

[Pea86] Judea Pearl. Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29:241–188, 1986.

[PGA87] D. Poole, R. Goebel, and R. Aleliunas. Theorist: A logical reasoning system for defaults and diagnosis. In N.J. Cercone and G. McCalla, editors, *The Knowledge Frontier: Essays in the Representation of Knowledge*, pages 331–352. Springer Verlag, New York, 1987.

[Poo88] D. Poole. A logical framework for default reasoning. *Artifical Intelligence*, 36(1):27–47, 1988.

[Prz89] Teodor C. Przymusinski. An algorithm to compute circumscription. *Artificial Intelligence*, 38(1):49–73, January 1989.

[Rei87] R. Reiter. Nonmonotonic reasoning. In *Annual Reviews of Computer Science*, pages 147–186. Annual Reviews of Computer Science, New York, 1987.

[RJ87] R. Reiter and De Kleer J. Foundations of assumption-based truth maintenance systems (preliminary report). In *Proceedings of AAAI-87*, pages 183–187, Seattle, Washington, July 13-17 1987.

[SG89] A. Sattar and R. Goebel. Using crucial literals to select better theories. Research report TR 89-27, Department of Computing Science, University of Alberta, Edmonton, Canada, October 1989.

Causal theories for default and abductive reasoning

Hector Geffner
hector@ibm.com
IBM T. J. Watson Research Center
P. O. Box 704
Yorktown Heights, NY 10598

1 Introduction

We explore the relation between default and abductive reasoning. Default reasoning is concerned with the adoption of assumptions in the absence of conflicting information, while abductive reasoning is concerned with the adoption of hypotheses that increase the coherence of a set of beliefs Formal accounts of default reasoning have encountered the problem of spurious arguments: arguments which rely on acceptable defaults but which support unacceptable conclusions. General accounts of abductive reasoning, on the other hand, have encountered the problems of identifying what needs to be explained, what counts as an explanation, and what hypotheses can be postulated.

The connection between default and abductive reasoning lies in the notion of expectation failures or "abnormalities." Default reasoning arises from the minimization of such abnormalities (e.g. [McCarthy, 1986]), while abductive reasoning arises from the need to explain them. However, not all "abnormalities" are equal. Many examples illustrate that in cases of conflict, one abnormality may be preferred to others (e.g. the famous "Yale shooting problem" [Hanks and McDermott, 1987]). Often the source of such preferences is related to the notion of *explanation*: some scenarios explain the abnormalities they introduce, while others do not. As expected, among the competing scenarios that arise from conflicting defaults, the most *coherent* ones usually capture the intuitive default expectations, while less coherent ones give rise to spurious expectations.

Causal theories are default theories which explicitly address the distinction between *explained* and *unexplained* abnormalities. They embed a "causal" operator 'C' in the language such that for an abnormality α, the literal $C\alpha$ is supposed to hold when α is *explained*. Such an operator permits to make explicit the causal or explanatory character of rules like "irregular ignition causes power decrease," "being sick explains being unable to go to class," and so on. The *interpretation* of causal theories makes use of such rules to identify most coherent scenarios and determine the conclusions which causal theories legitimately support.

Causal theories have been analyzed in detail in [Geffner, 1989a] where they are shown to provide a *unifying framework* for several domains of interest to AI, including temporal and abductive reasoning, and logic programs with negation (see also [Geffner, 1989b] for related ideas). In this extended abstract I will summarize the main concepts and results, with special emphasis on the application of causal theories to *abductive reasoning*. Due to the distinction between *explained* and *unexplained* expectation failures, causal theories not only eliminate spurious default arguments, but also permit to identify what needs to be explained, what counts as an explanation, and what hypotheses increase the coherence of a set of beliefs.

2 Causal Theories

A causal theory is a classical first order theory augmented with the causal operator 'C', in which certain atoms, *abnormalities*, are expected to be false [McCarthy, 1986]. We use the symbol α possibly indexed to denote abnormalities. Furthermore, for an interpretation M, $A[M]$, called the *gap* of M, denotes the set of abnormalities true under M. As usual, an interpretation M that satisfies a causal theory T is said to be a model of T. A *class* $\mathcal{C}$ of T with a *gap* $A[\mathcal{C}]$ stands for the non-empty collection of models M of T with a gap $A[M] \subseteq A[\mathcal{C}]$. Intuitively, since the negation of abnormalities are *assumptions* expected to hold, a class $\mathcal{C}$ with a gap $A[\mathcal{C}]$ represents the collection of models which validate all assumptions $\neg\alpha$ for abnormalities α not in $A[\mathcal{C}]$. We say that a proposition p *holds* in a class $\mathcal{C}$ of T, if p holds in every model in $\mathcal{C}$. Proof-theoretically this is equivalent to require that p be logically derivable from T and a set of assumptions compatible with $A[\mathcal{C}]$.

The operator C is most commonly used to encode causal or explanatory rules of the form "if a then b" as sentences of the form $a \Rightarrow Cb$ (see [Pearl, 1988]). A rule such as "rain causes the grass to be wet" may thus be expressed as $\texttt{rain} \Rightarrow C\texttt{grass\_wet}$, which can be read as saying that if **rain** is true, then **grass_wet** is explained. The operator C obeys certain constraints; here, for simplicity, we will only require α to hold when

$C\alpha$ holds (namely, α must be true when α is explained).

The operator C induces a preference relation on *classes* of models of the theory T of interest, which is used to determine the propositions (causally) *entailed* by T (see [Shoham, 1988], for this non-classical form of entailment). Such preference relation depends on the abnormalities and the *explained* abnormalities sanctioned by the different classes. An abnormality α is *explained* in a class C when the literal $C\alpha$ holds in C. If we denote by $A^e[C]$ the set of explained abnormalities in a class C of a theory T, then the preference relation on classes can be described as follows.

A class C is *as preferred as* a class C' iff $A[C] - A^e[C] \subseteq A[C']$. A class C is *preferred to* a class C' iff C is as preferred as C' but C' is *not* as preferred as C.

In other words, a class C is preferred to a class C' when every abnormality in C but not in C' has an explanation, but not vice versa. If there is no class preferred to C, then C is said to be a *preferred class*. A causal theory T *(causally) entails* a proposition p if p holds in all its preferred classes.

It is simple to show that in order to determine the propositions which are entailed by a theory T it is sufficient to consider the *minimal* classes of T; namely, those classes C of T with a minimal gap $A[C]$.

Example 1 Let us consider first a simple causal theory T given by the single sentence $\neg\mathbf{ab}_1 \Rightarrow C\mathbf{ab}_2$, where $\mathbf{ab}_1$ and $\mathbf{ab}_2$ are two different abnormalities. Such a theory admits two minimal classes: a class C_1, comprised by the models of T which only sanction the abnormality $\mathbf{ab}_1$, and a class C_2, comprised of the models which only sanction the abnormality $\mathbf{ab}_2$. Thus C_1 has an associated gap $A[C_1] = \{\mathbf{ab}_1\}$, while C_2 has an associated gap $A[C_2] = \{\mathbf{ab}_2\}$. Both classes represent the minimal classes of T, as there is no model of T that satisfies both assumptions $\neg\mathbf{ab}_1$ and $\neg\mathbf{ab}_2$, together with the restriction $C\alpha \Rightarrow \alpha$. The abnormalities α explained in each class C can be determined by testing which literals $C\alpha$ hold in C. As we said, this amounts checking whether there is a set of assumptions legitimized in C which together with T logically imply $C\alpha$. In the class C_2, hence, the abnormality $\mathbf{ab}_2$ is explained as the literal $C\mathbf{ab}_2$ logically follows from T and the assumption $\neg\mathbf{ab}_1$ legitimized in C_2. On the other hand, the abnormality $\mathbf{ab}_1$ is *not* explained in C_1, as there is no set of assumptions validated by C_1 which supports the literal $C\mathbf{ab}_1$. It follows then, that the class C_2 is causally preferred to C_1, as $A[C_2] - A^e[C_2] = \emptyset \subseteq A[C_1]$, but $A[C_1] - A^e[C_1] = \{\mathbf{ab}_1\} \not\subseteq A[C_2] = \{\mathbf{ab}_2\}$. Furthermore, since C_1 and C_2 are the only minimal classes of T, C_2 remains as the single causally preferred class of T, and the propositions $\neg\mathbf{ab}_1$ and $\mathbf{ab}_2$ are (causally) entailed by T.

Note the asymmetry established by the causal operator in the sentence $\neg\mathbf{ab}_1 \Rightarrow C\mathbf{ab}_2$; while the assumptions $\neg\mathbf{ab}_1$ and $\neg\mathbf{ab}_2$ are incompatible, the former is legitimized while the latter is not.

Example 2 (Reasoning about change) Let us consider now a bare bones description T of the Yale shooting problem [Hanks and McDermott, 1987] :

(1) $\mathtt{loaded}_0 \wedge \neg\mathbf{ab}_1 \Rightarrow \mathtt{loaded}_1$

(2) $\mathtt{alive}_1 \wedge \neg\mathbf{ab}_2 \Rightarrow \mathtt{alive}_2$

(3) $\mathtt{shoot}_1 \wedge \mathtt{loaded}_1 \Rightarrow C\neg\mathtt{alive}_2$

(4) $\mathtt{shoot}_1 \wedge \mathtt{loaded}_1 \Rightarrow C\mathbf{ab}_2$

(5) $\mathtt{loaded}_0 \wedge \mathtt{alive}_1 \wedge \mathtt{shoot}_1$

This simplified version preserves the main features of the original problem and gives rise to the same anomalies pointed out by Hanks and McDermott. The causal operator, however, makes now the causal character of rules (3) and (4) explicit. In the context of T, the assumption $\neg\mathbf{ab}_1$ about the persistence of **loaded** is in conflict with the assumption $\neg\mathbf{ab}_2$ about the persistence of **alive**, and thus two minimal classes of models arise: a class C_1 of models in which the former abnormality holds, and a class C_2 of models in which the later abnormality hold. However, since the assumption $\neg\mathbf{ab}_1$ *explains* the abnormality $\mathbf{ab}_2$ in T, i.e. $T, \neg\mathbf{ab}_1 \vdash C\mathbf{ab}_2$, but not vice versa, the latter class of models is preferred. As a result, the expected literals $\mathtt{loaded}_1$ and $\neg\mathtt{alive}_2$ are (causally) entailed by T.

Example 3 (Logic Programs with Negation) Consider a logic program P given by the rules:

$$\mathbf{c} \leftarrow \mathbf{a}, \neg\mathbf{b}$$

$$\mathbf{d} \leftarrow \neg\mathbf{c}$$

$$\mathbf{a} \leftarrow$$

and the causal theory $C[P]$:

$$C\mathbf{a} \wedge \neg\mathbf{b} \Rightarrow C\mathbf{c}$$

$$\neg\mathbf{c} \Rightarrow C\mathbf{d}$$

$$\mathbf{t} \Rightarrow C\mathbf{a}$$

obtained by translating each rule

$$\gamma \leftarrow \alpha_1, \ldots, \alpha_n, \neg\beta_1, \ldots, \neg\beta_m$$

in P, into a *causal* rule:

$$C\alpha_1 \wedge \ldots \wedge C\alpha_n \wedge \neg\beta_1 \wedge \ldots \wedge \neg\beta_m \Rightarrow C\gamma \quad .$$

Provided that all *non-causal* atoms (atoms which do not involve the operator C) are considered 'abnormal,' it can be shown that the canonical semantics of P [Apt *et al.*, 1987] and the interpretation of the causal theory $C[P]$ legitimize the same behavior. Moreover, the same equivalence holds *for any stratified program* P [Geffner, 1989a]. In other words, causal theories provide an alternative semantics for characterizing logic programs with negation.

3 Abductive Reasoning

Having illustrated the generality of causal theories, we focus now on their application to abductive reasoning [Peirce, 1955, Charniak and McDermott, 1985]. The

central idea is to associate a *coherence* measure to *contexts* (theories) as opposed to *classes* of models, and to identify abduction with the adoption of hypotheses which render a theory more coherent. Intuitively, the coherence —or, for that matter, the incoherence— of a context T will depend on whether the abnormalities it declares are explained.

Formally, if $\mathcal{C}_i$, $i = 1, \ldots, n$ are the preferred classes of T, we define the *incoherence set* $I[T]$ of T to be the collection of sets $A^u[\mathcal{C}_i] = A[\mathcal{C}_i] - A^e[\mathcal{C}_i]$, $i = 1, \ldots, n$, where $A[\mathcal{C}_i]$ and $A^e[\mathcal{C}_i]$ stand for the gap and the explained gap of class $\mathcal{C}_i$ respectively.

A context T is *as coherent as* as a context T' then, if for every set S in $I[T]$ there is a set S' in $I[T']$ such that $S \subseteq S'$. Furthermore, T is *more coherent* than T' if T is as coherent as T', but T' is not as coherent as T.

For example, a context in which Tim is known to be a *sick* non-flying bird, is *more coherent* than a context in which all that is known is that Tweety is a non-flying bird, as the former provides an explanation for Tim's unexpected feature.

Given a pool Ξ of possible conjectures that can be adopted as hypotheses (see [Poole, 1987]), we define *belief states* as contexts $T + \Xi$ that result from augmenting a causal theory T with a set of conjectures Ξ, $\Xi \in \Xi$, logically consistent with T. We often denote such states as pairs $\langle T, \Xi \rangle$ to distinguish the solid evidence T from the hypothetical beliefs Ξ. The definitions below will permit us to derive Ξ from T in such a way that the resulting theory $T + \Xi$ is *maximally coherent* and devoid of unnecessary commitments.

First, we say that a belief state $\mathcal{B} = \langle T, \Xi \rangle$ is *less committed* than a belief state $\mathcal{B}' = \langle T, \Xi' \rangle$, if $\Xi \subset \Xi'$. Likewise, we say that $\mathcal{B} = \langle T, \Xi \rangle$ is a *maximally-coherent* belief state if there is no other belief state $\mathcal{B}' = \langle T, \Xi' \rangle$ more coherent than $\mathcal{B}$.

Finally, a maximally-coherent belief state $\mathcal{B} = \langle T, \Xi \rangle$ is *admissible*, when there is no maximally-coherent belief state $\mathcal{B}' = \langle T, \Xi' \rangle$ less committed than $\mathcal{B}$. Intuitively, admissible belief states $\langle T, \Xi \rangle$, represent belief states which a rational agent with the information in T may choose to adopt. We also say in that case that Ξ is an *admissible hypothesis set* in T, and that the conjectures in Ξ are *admissible hypotheses*.

Example 4 (Simple Abduction) Consider the causal network depicted in fig. 4 describing a fragment of the knowledge relevant to the diagnosis of a malfunctioning car [Console *et al.*, 1989]. We encode such a network by mapping each causal link $\alpha \rightarrow \beta$ to a causal rule $\alpha \Rightarrow C\beta$, and by regarding each token in the net as an 'abnormality' that needs explanation. Furthermore, we assume a pool Ξ of conjectures which includes only the top propositions **pistons_rings_used**, **oil_cup_holed**, **old_spark_plugs**, all of which are assumed to be self-explanatory; namely, for each such conjecture ξ we assume $\xi \Rightarrow C\xi$.

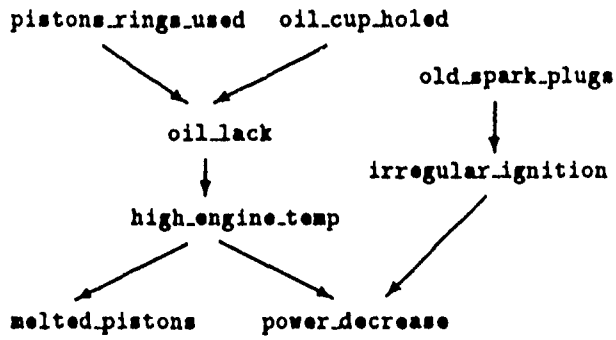


Figure 1: A causal network

Let us assume now that **power_decrease** is observed, and let T refer to the resulting context. Such context accepts a single preferred class of models in which the abnormality **power_decrease** holds but is not explained. Any belief state $\mathcal{B}_i = \langle T, \Xi_i \rangle$ with a non-empty set of conjectures Ξ_i will explain such incoherence away, and indeed, any such belief state will be *maximally-coherent*. However only those states containing a *single* hypothesis from Ξ will qualify as *admissible* belief states. Thus, there are three admissible belief states, involving three singleton hypothesis sets {**pistons_rings_used**}, {**oil_cup_holed**}, and {**old_spark_plugs**} respectively. If ¬**high_engine_temp** is further observed, however, only one admissible hypothesis would remain: **old_spark_plugs**; both **pistons_rings_used** and **oil_cup_holed** require **high_engine_temp** in order to explain **power_decrease**. Note that if ¬**irregular_ignition** also becomes available, no admissible hypothesis would remain. In such context, there would be a single (incoherent) belief state involving no conjecture at all. Such behavior is thus different from the behavior sanctioned by approaches in which abduction is viewed as deduction in a *completed model* (e.g. [Kautz, 1987] and [Console *et al.*, 1989]).

It is common to find two different types of diagnostic tasks in the AI literature: *abductive diagnosis*, in which the search is for hypotheses that imply the observations, and *consistency-based diagnosis*, in which the search is for hypotheses that render the model and the observations consistent (see [Poole, 1989]). The example above, for instance, belongs to the first category. There is, however, a natural way in which consistency-based diagnosis can also be accommodated within the present framework. All that is needed is to stipulate that 'abnormalities' are self-explanatory, i.e., we need to assert expressions of the form $\alpha \Rightarrow C\alpha$ for every relevant 'abnormality' α. In that case all minimal classes will be perfectly coherent, and thus, no need for abductive hypotheses would ever arise.

There are however other patterns of abductive inference which cannot be accommodated so easily. For instance, we may wish to express above that the hypothesis **pistons_rings_used** is *more likely* than the hypothesis **oil_cup_holed**. For that we not only need to be able to specify the pool of conjectures, but also how such conjectures are supposed to be ordered. The extension below addresses this limitation and shows how this additional information can be used to prune the space of admissible hypotheses.

A *preference relation on conjectures* is a strict partial order on the set Ξ of conjectures. We denote such an ordering by the symbol '$\succ$.' The expression $\xi \succ \xi'$ is thus to be read as stating that conjecture ξ is preferred to conjecture ξ'. The preference relation is extended to sets of conjectures as follows.

A *set of conjectures* Ξ *is preferred to a set of conjectures* Ξ', if every conjecture in $\Xi - \Xi'$ is preferred to some conjecture in $\Xi' - \Xi$.[1] Similarly, a *maximally-coherent* belief state $\mathcal{B} = \langle T, \Xi \rangle$ is a *preferred belief state* in context T, if there is no other *maximally-coherent* belief state $\mathcal{B}' = \langle T, \Xi' \rangle$ with an hypothesis set Ξ' preferred to Ξ.

The pair formed by a causal theory and an ordered set of conjectures constitute an *abductive causal theory*. We illustrate the application of abductive causal theories in a diagnostic task of the type considered by Reggia *et al.* [1985].

Example 5 Let us consider the causal network shown in fig. 5, We assume that $\mathtt{d}_i$'s, $i = 1, \ldots, 5$ stand for diseases, and that $\mathtt{m}_j$'s, $j = 1, \ldots, 3$ stand for manifestations. Each link of the form $\alpha \rightarrow \beta$ is expressed as a causal rule of the form $\alpha \Rightarrow C\beta$, except the link $\mathtt{d}_1 \rightarrow \mathtt{m}_1$, modified by $\mathtt{d}_5$, which is mapped into the rule $\mathtt{d}_1 \wedge \neg \mathtt{s}_1 \Rightarrow C\mathtt{m}_1$, where $\mathtt{s}_1$ is an exceptional state which can be explained by $\mathtt{d}_5$, i.e. $\mathtt{d}_5 \Rightarrow C\mathtt{s}_1$. The manifestations $\mathtt{m}_j$, $j \in [1,3]$, and the state s_1 constitute the space of 'abnormalities'.[2] The diseases, $\mathtt{d}_i$, $i \in [1,5]$, on the other hand, represent an ordered set of conjectures such that $\mathtt{d}_i$ is preferred to $\mathtt{d}_j$ iff $i > j$. Such preferences may be available for instance from the corresponding prior probabilities.

Consider now that $\mathtt{m}_2$ is observed. From the network depicted in fig. 5, it is simple to see that $\mathtt{m}_2$ gives rise to three admissible belief states $\mathcal{B}_i = \langle \{\mathtt{m}_2\}, \{\mathtt{d}_i\} \rangle$, for $i = 1, 2, 3$. However, due to the preference order on hypotheses, only the belief state $\mathcal{B}_1$ and the hypothesis $\mathtt{d}_1$ remain preferred.

Let us further assume that, refuting the hypothesis $\mathtt{d}_1$, $\neg\mathtt{m}_1$ is now observed. This new observation leads again to three admissible belief states; however, while conjectures $\mathtt{d}_2$ and $\mathtt{d}_3$ remain admissible singleton hypotheses sets in new context, the third hypothesis set is now given by the compound hypothesis $\{\mathtt{d}_1, \mathtt{d}_5\}$. Furthermore, due to the preference order on conjectures, the hypothesis $\mathtt{d}_2$ becomes the single preferred hypothesis, as is preferred to $\mathtt{d}_3$ and $\mathtt{d}_5$.

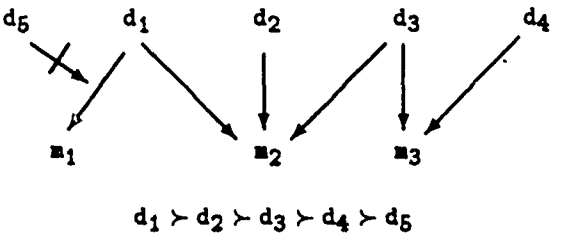


Figure 2: A simple diagnostic model

Finally, let us assume that $\mathtt{m}_3$ is observed. The new context gives rise again to three admissible belief states, involving the admissible hypotheses sets $\{\mathtt{d}_1, \mathtt{d}_4, \mathtt{d}_5\}$, $\{\mathtt{d}_2, \mathtt{d}_4\}$, and $\{\mathtt{d}_3\}$ respectively. However, due to the preferences on conjectures, this time $\mathtt{d}_3$ remains as the single leading hypothesis, followed by the compound hypotheses $\{\mathtt{d}_2, \mathtt{d}_4\}$, and only then by $\{\mathtt{d}_1, \mathtt{d}_2, \mathtt{d}_5\}$.

Acknowledgments

I want to thank Judea Pearl whose work on evidential reasoning shaped my intuitions on this and related topics.

References

[Apt *et al.*, 1987] K. Apt, H. Blair, and A. Walker. Towards a theory of declarative knowledge. In J. Minker, editor, *Foundations of Deductive Databases and Logic Programming*, pages 89–148. Morgan Kaufmann, Los Altos, CA, 1987.

[Charniak and McDermott, 1985] E. Charniak and D. McDermott. *Introduction to Artificial Intelligence.* Addison Wesley, Reading, MA., 1985.

[Console *et al.*, 1989] L. Console, D. Dupre, and P. Torasso. A theory of diagnosis for incomplete causal models. In *Proceedings IJCAI-89*, pages 1311–1317, Detroit, Michigan, 1989.

[Geffner, 1989a] H. Geffner. *Default Reasoning: Causal and Conditional Theories.* PhD thesis, UCLA, Los Angeles, CA, November 1989.

[Geffner, 1989b] H. Geffner. Default reasoning, minimality and coherence. In *Proceedings of the First International Conference on Principle of Knowledge Representation and Reasoning*, pages 137–148, Toronto, Ontario, 1989.

[1] This preference relation is still a simplification. Additional factors like the cardinality of the hypotheses sets Ξ and Ξ' are likely to play a role in the preference of Ξ over Ξ'.

[2] The same results would follow if diseases were treated as abnormalities, provided that they are self-explanatory.

[Hanks and McDermott, 1987] S. Hanks and D. McDermott. Non-monotonic logics and temporal projection. *Artificial Intelligence*, 33:379–412, 1987.

[Kautz, 1987] H. Kautz. *A Formal Theory of Plan Recognition.* PhD thesis, University of Rochester, Rochester, N.Y., May 1987 .

[McCarthy, 1986] J. McCarthy. Applications of circumscription to formalizing commonsense knowledge. *Artificial Intelligence*, 28:89–116, 1986.

[Pearl, 1988] J. Pearl. Embracing causality in default reasoning. *Artificial Intelligence*, 35:259–271, 1988.

[Peirce, 1955] C. Peirce. *Abduction and Induction.* Dover, New York, 1955.

[Poole, 1987] D. Poole. Defaults and conjectures: hypothetical reasoning for explanation and prediction. Technical Report CS-87-4, University of Waterloo, 1987.

[Poole, 1989] D. Poole. Normality and faults in logic-based diagnosis. In *Proceedings IJCAI-89*, pages 1304–1310, Detroit, Michigan, 1989.

[Reggia *et al.*, 1985] J. Reggia, D. Nau, P. Wang, and Y. Peng. A formal model of abductive inference. *Information Sciences*, 37:227–285, 1985.

[Shoham, 1988] Y. Shoham. *Reasoning about Change: Time and Causation from the Standpoint of Artificial Intelligence.* MIT Press, Cambridge, Mass., 1988.

Probabilistic and Qualitative Abduction *

Judea Pearl

Computer Science Department
University of California
Los Angeles, California 90024

Abstract

This paper discusses relations between the probabilistic and qualitative approaches to abduction; it then offers a probabilistic account of the connection between causation and explanation, and proposes a non-temporal probabilistic semantics to causality.

1. Introduction

In the probabilistic approach, abduction is considered the task of finding the "most probable explanation" of the evidence observed, namely, seeking an instantiation of a set of explanatory variables that attains the highest probability, conditioned on the evidence observed. The qualitative approaches make explicit appeal to explanatory scenarios, and seek scenarios that are both coherent and parsimonious.

The major challenge for both the probabilistic and the qualitative approaches is to enforce an appropriate separation between the *prospective* and *retrospective* modes of reasoning so as to capture the intuition that prediction should not trigger suggestion. To use my favorite example: "Sprinkler On" predicts "Wet Grass," "Wet Grass" suggests "Rain," but "Sprinkler On" should not suggest "Rain." In the probabilistic approach such separation is enforced via patterns of independencies that are assumed to accompany causal relationships, cast in conditional probability judgments. In the qualitative approaches the separation is accomplished in two ways. One is to label sentences as either *causally established* (i.e., explained) or *evidentially established* (i.e., conjectured) and subject each type to a different set of inference rules [Pearl 1988a; Pearl 1988b; Geffner 1989]. The second method is to regard abduction as a specialized meta-process that operates on a causal theory [Poole 1987; Reiter 1987].

The obvious weakness of the qualitative approach is the lack of rating among competing explanations and, closely related to it, the lack of ratings of pending information sources. On the other hand, qualitative strategies demand fewer judgments in constructing the knowledge base.

In qualitative theories simplicity is enforced by explicitly encoding the preference of simple theories over complex ones, where simple and complex are given syntactical definitions, e.g., smallest number of (cohesive) propositions [Thagard 1989], minimal covering [Reiter 1987; Reggia et al. 1983]. These syntactic ratings do not always coincide with the notion of plausibility, for example, two common diseases are often more plausible than a single rare disease in explaining a given set of symptoms [Reggia 1989]. In probabilistic theories, coherence and simplicity are managed together by one basic principle — maximum posterior probability.

2. Explanation and Causation

Explanations are intimately connected to causation. When we say that "a explains b" we invariably assume the existence of a causal theory according to which "a tends to cause b" and, furthermore, that in the particular situation where b was observed, "a actually caused b." The subtle difference between "tends to explain" and "actually caused" has been the subject of much discussion in the philosophical literature, a summary of which can be found in

* This work was supported in part by National Science Foundation Grant #IRI-88-21444 and Naval Research Laboratory Grant #N00014-89-J-2007.

Skyrms & Harper [1988]. The classical example amplifying this difference is that of a skillful golfer who makes a shot with the intention of getting the ball in the hole; the shot is actually quite poor, but the ball hits a tree branch and is deflected into the hole. Here, we are likely to say that the golfer's skill and attention "tended to cause." but did not "actually cause" the ball to get in the hole. Explanation is connected with the latter, not the former; the phrase "tends to explain" is hardly in use in the language, instead, we use the phrase "is normally suggested by."

In the language of probability this distinction can be related to a difference between two conditional probabilities. If C has a tendency to cause E, then we expect $P(E|C)$ to be high. If C is identified as the event that "actually caused" E, then we expect $P(C|E, context)$ to be high where, by *context*, we mean other facts connected with the observation of E (e.g., hitting the tree in the golfer example).

In general, the probability $P(E|C)$ stands for a mental summary of a vast number of scenarios leading from C to E. Some of these scenarios involve contingencies such as trees intercepting golf balls, and some involve micro processes that can be articulated only at more refined levels of abstraction, for example, the interactions between the golf ball and the ground particles. When we confirm the sentence "C actually caused E" we normally mean that some path of contiguous micro events either can be presumed to have taken place or was actually observed. Such events are encoded in a knowledge strata more refined than the one used in the main discourse. For example, a pathologist may assert that the bullet was the "actual" cause of death only if a collection of key anatomical findings are observed confirming the existence of a contiguous physiological process leading from the bullet entry to death.

3. What's in an Explanation, a Probabilistic Proposal

If abduction is defined as "inference to the best explanation", a natural question to ask is how we define an explanation. Both the probabilistic and qualitative approaches to abduction have so far treated the term "explain" as a given primitive relationship among events, from which a "best" overall explanation is to be assembled. Both approaches have given the term "explain" a procedural semantics, attempting to match the way people use it in inference tasks, but were not concerned with what makes people believe that "a explains b," as opposed to, say, "b explains a" or "c explains both a and b." The quest for an empirical semantics of explanation has a long history in the literature of probabilistic causality, where the focus has been finding an operational definition of causation. (see Reichenbach [1956]; Simon [1957]; Good [1961]; Salmon [1984]; Suppes [1970]; Glymour et al. [1987]; Skyrms [1988]).

ith the exception of Simon [1957] and Glymour et al. [1987], temporal precedence was assumed to be essential for defining causation. For example, Reichenbach (1956, page 204) says that C is *causally relevant* to E if:

(i) $P(E|C) > P(E)$

(ii) There is no set of events *earlier* than, or simultaneous with, C such that conditional on these events E and C are probabilistically independent.

Suppes [1970] subscribes to a similar definition, with an explicit requirement that C precedes E in time.

These criteria offer a working definition for causation provided that the observed dependencies are not produced by *hidden* causes and provided that the set of events mentioned in condition (ii) is restricted to be "natural" events, excluding *artificial* events, syntactically concocted to meet condition (ii) [Good 1961; Suppes 1984].

I would like now to propose a non-temporal extension of the Reichenbach-Suppes definition of causation, one that determines the direction of causal influences without resorting to temporal information. It should be applicable, therefore, to the organization of concurrent events or events whose chronological precedence cannot be determined empirically. Such situations are common in the behavioral and medical sciences where we say, for example, that old age explains a certain disability, not the other way around, even though the two occur together (in many cases it is the disability that precedes old age). Similarly, we say that an incoming rain storm explains the falling barometer although, perceptually, the latter precedes the former in time.

The intuition behind my definition revolves around the perception of *voluntary control* [Simon 1980] and its probabilistic formulation in terms of conditional independence (see Pearl [1988], page 396). The reason we insist that the

rain caused the grass to become wet and not that the wet grass caused the rain is that we can create conditions which, without disrupting the natural dependence between rain and wet grass, can get the grass wet without affecting the rain. We can, of course, also create a situation where the rain falls and the grass remain dry, say by seeding the clouds and covering the grass, but under such conditions the dependence between rain and wet grass is disrupted, which violates the symmetry between the two procedures.

As was stressed in Pearl [1988, page 396], the perception of voluntary control is not a necessary element in this asymmetry between cause and effect, but may in itself be a bi-product of dependencies observed among uncontrolled variables. In medical research, for example, we often search for a causal culprit of a disease much before attaining control over such cause.

Articulating these considerations in probabilistic terms, we come up with the following non-temporal extension of the Reichenbach-Suppes definition.

Definition: (non-temporal causation) An event C is said to be a (tentative) *direct cause* of E if

(i) $P(E|C) > P(E)$

(ii) There is no set of events such that conditional on these events E and C are independent.

(iii) There is an event C' and a set S of events not containing C, E and C' such that:

$$P(E|S,C') > P(E|S), \text{ and}$$

$$P(C|S,C') = P(C|S)$$

The set S in (iii) represents conditions needed for eliminating possible spurious dependencies between C and C'. Event C' represents our means for gaining control over E, namely, an event that can cause E without affecting C, thus providing an alternative explanation to E. Ironically, and almost circularly, explanations are defined in terms of their very destruction by other explanations; C qualifies as an explanation of E only if it can be "explained away" or rendered superfluous by some alternative explanation of C'. This is not surprising in view of the fact that people often seek an explanation for the sole purpose of ruling out others. For example, I often hope that my broker would explain the falling prices of my stock in terms of investors' panic and other transitory phenomena, so as to allay my fears of more profound explanations.

Any non-temporal definition of causation immediately raises the question of consistency, for example, is it possible that using criteria (i) through (iii) we would generate two incompatible assertions: "C cause E" and "E causes C?" It can be shown, however, that for a larger class of probability distributions these criteria are safe from such inconsistencies. Moreover, for those distributions that are unsafe, we can constrain (iii) by an additional restriction:

(iv) For every set of events S' that does not contain E and C, if there is an event E' (not in S') such that

$$P(C|S',E') > P(C|S'),$$

then

$$P(E|S',E') \neq P(E|S').$$

This restriction guarantees that we certify C as a direct cause of E only if the criterion (iii) is violated when we interchange C and E.

The definition above is a translation of that given in Pearl [1988b] to the language of Reichenbach and Suppes, where causes are propositional events having "positive" influence, hence the inequality in (i). In Pearl [1988b] these conditions were articulated in terms of *variables* rather than positively influencing *events*. A similar definition, in terms of variables, was introduced in Spirtes et al. [1989].

Another variant of this definition can be articulated using the graphical language of Bayesian networks, by considering all $n!$ orderings in which such a network can be constructed. We say that a variable C is a *direct cause* of variable E if:

(1) C and E are adjacent in all orderings, and

(2) There is an ordering in which C is a free parent of E, i.e., non-adjacent to some other parent of E, and there is no ordering in which E is a free parent of C.

This formulation reveals the type of empirical asymmetry that is responsible for evoking the perception of directionality in causal relationships.

On the practical side we also must address the question of

computation complexity since, in principle, conditions (ii) and (iii) call for testing all subsets of events. It can be shown that, for a larger class of probability distributions, effective algorithms exist that determine the direction of causal influences without testing all subsets of events [Geiger 1990; Verma 1990].

A question of a more philosophical flavor concerns the relation between temporal precedence and the orientations determined by our definition: Why is it that we never observe a clash between the two? The answer, I believe, lies in the flexibility of our language; whenever the flow of dependency-based causality seems to clash with the direction of time we invent new variables (hidden causes) that reverse the former to comply with the latter.

References

Geffner, H. 1989. Default reasoning: Causal and conditional theories. Phd. dissertation. Cognitive Systems Laboratory *Technical Report (R-137)*. Computer Science Dept. University of California, Los Angeles.

Geiger, D. 1990. Graphoids: A qualitative framework for probabilistic inference. Phd. dissertation. Cognitive Systems Laboratory *Technical Report (R-142)*. Computer Science Dept. University of California, Los Angeles.

Glymour, C., Scheines, R., Spirtes, P., and Kelly, K. 1987. *Discovering causal structure*. New York: Academic Press.

Good, I.J. 1961. A causal calculus. *British Journal for Philosophy of Science* Vol. 11:305-328; 12, 43-51; 13, 88; reprinted as Ch. 21 in: *Good Thinking* (University of Minnesota Press, Minneapolis, MN, 1983).

Pearl, J. 1988a. Embracing causality in formal reasoning. *Artificial Intelligence* 35(2):259-71.

Pearl, J. 1988b. *Probabilistic reasoning in intelligent systems*. San Mateo: Morgan Kaufmann.

Poole, D. L. 1987. Defaults and conjectures: Hypothetical reasoning for explanation and prediction. Research Report CS-87-54, University of Waterloo (Kitchener, Ontario).

Reggia, J.A. 1989. Measuring the plausibility of explanatory hypotheses. *Behavioral and Brain Sciences* Vol. 12(3): 485-487.

Reggia, J. A., Nau, D. S., and Wang, Y. 1983. Diagnostic expert systems based on a set-covering model. *Intl. Journal of Man-Machine Studies* 19: 437-60.

Reichenbach, H. 1956. The direction of time. Berkeley, CA: University of California Press.

Reiter, R. 1987. A theory of diagnosis from first principles. *Artificial Intelligence* 32(1): 57-95.

Salmon, W. 1984. Scientific explanation and the causal structure of the world. Princeton: Princeton University Press.

Simon, H. 1957. *Models of man*. New York: Wiley and Sons.

Simon, H.A. 1980. The meaning of causal ordering. In *Qualitative and quantitative social research*, ed. R. K. Merton, J. S. Coleman, and P. H. Rossi, 65-81. New York: Free Press.

Skyrms, B. 1988. Probability and causation. *Journal of Economics* Vol. 39, 53-68.

Skyrms, B. and Harper, W.L. 1988. *Causation, chance and credence*. Dordrecht: Kluwer Academic Publisher.

Spirtes, P. 1989. Causality from probability. Dept. of Philosophy, Carnegie-Mellon University, Report #CMU-LCL-89-4.

Suppes, P. 1970. *A probabilistic theory of causality*. Amsterdam: North Holland.

Suppes, P. 1984. *Probabilistic metaphysics* Oxford: Blackwell.

Thagard, P. 1989. Explanatory coherence. *Behavioral and Brain Sciences* Vol. 12(3): 435-468.

Verma, T. 1990. Learning causal structure from independence information. (in preparation).